

# NBN Record Cleaner user guide

Version 1.0.8.3

August 2012  
Stuart Ball, JNCC  
Graham French, NBN Trust

## Contents

1. Purpose .....	3	6. Allocating Vice-counties .....	24
1.1. Validation .....	3	6.1. How VCs are allocated.....	25
1.2. Verification .....	3	7. Verification.....	26
1.3. Summary.....	4	7.1. Identification difficulty.....	27
2. Installation and updates.....	5	7.2. Mapping.....	27
2.1. Requirements .....	5	7.3. Filtering.....	30
2.2. Record Cleaner installation.....	5	7.4. Running verification rules .....	30
2.3. Software updates .....	5	7.5. Saving results .....	32
2.4. Obtaining verification rules and keeping them up to date .....	6	8. Managing verification rule files .....	33
2.5. Limitations.....	6	8.1. How rules are supplied .....	33
3. Basic use of the Record Cleaner .....	6	8.2. How rules are updated.....	33
4. Preparing and loading data.....	7	8.3. Choosing rule-sets.....	34
4.2. Loading your data .....	8	9. Loading data from other databases .....	35
4.3. Matching columns.....	13	10. Use with MapMate .....	37
4.4. Including additional fields.....	19	11. Use with Recorder 6 .....	38
4.5. Saving a template .....	19	12. Use with NBN Exchange Format .....	39
5. Validation .....	21	13. Appendix.....	42
5.1. Fixing errors.....	22	13.1. Installation directories.....	42
5.2. Saving failed rows and rows you have edited .....	23	13.2. Data loading and validation error messages .....	43
5.3. Fixing the original data file .....	23	13.3. Reinstalling on Windows 7 machine	46
5.4. Proceed without fixing errors.....	24	13.4. Clearing verification rules.....	47

## 1. Purpose

The NBN Record Cleaner is designed to assist individual recorders, and organizations such as Local Record Centres and Recording Schemes, to check their data in a consistent and straightforward way in order to spot common problems. This should aid the process of data cleaning and ensure the quality of datasets passed on to others.

It is designed to be able to access biological record data stored in a wide variety of formats including various types of text files, Excel spreadsheets and databases - including records stored in the databases managed by biological recording packages such as *Recorder* and MapMate.

The checks are divided into two phases, termed “validation” and “verification”.

### 1.1. Validation

*“validation is the process of checking if something satisfies a certain criterion”*

Validation applies a built-in set of rules to the input data and reports cases where it does not pass one of these checks, including a brief explanation (e.g. “Invalid date”, “Species name not recognized”).  
**In such cases, the user should edit the original data to correct the problems and then retest.**

In our context, validation is the process of checking that the data presented can be interpreted successfully. For example:

- Is the date valid? “25/02/84” can be interpreted as a valid date (25<sup>th</sup> February 1984), but “31/02/84” cannot because 31<sup>st</sup> February does not exist!
- Can the spatial reference be recognized? “TL 1234” can be interpreted as an OSGB 1km square grid reference, but “TL12345” cannot because grid references must have an even number of digits.
- Are the names of species and vice counties recognisable? These are checked against list of acceptable terms: the Master Species list for species (which translates scientific and common names to TAXON\_VERSION\_KEYS) or the Vice County list (which translates names and common codes or abbreviations to vice-county numbers).

Rows that fail validation checks are excluded from the verification phase.

### 1.2. Verification

*“confirmation: additional proof that something that was believed (some fact or hypothesis or theory) is correct”*

During the verification phase, rules selected by the user are applied and cases are reported where the recorded value falls outside the acceptable range defined by a chosen rule.  
**Such cases must be assessed by the user when deciding what action (if any) to take.**

In our context, verification checks whether the data is credible and gives the user warning if it is unusual in some way. For example:

- Is the spatial reference given for a terrestrial animal actually on land? (or, conversely, is a record of a marine organism not on the land?)
- If both a spatial reference and a vice county are given, is the spatial reference within the stated vice county's boundary?
- Is the date given for an observation credible in relation to the biology of the species? A record of a Swallow in November is likely to raise eyebrows and require a higher standard of proof than a record from the same locality in July.

These checks are not built into the application but require additional information that is supplied in the form of rule files. For example, one of these rules might state the earliest and latest dates in the year on which records of Swallow are accepted without further question. Such externally defined rules codify expert knowledge about species and are supplied by expert groups such as National Recording Schemes and Societies. Record Cleaner provides an automated way of acquiring these rules and keeping them up to date via the Internet.

For example, if a record of a terrestrial species is not on land or a grid reference is not within the stated vice county, this is very like a validation failure and requires checking the original data to discover and correct the error. However, the case of a record of Swallow in November is a little different. This could be an error in the data (e.g. the wrong species name was entered); it could be due to mistaken identification by the observer; or it might actually be correct (odd records of migrant birds out of season are not unknown!). This will usually require going back to the observer to check whether the data was correctly entered and, if so, whether they have additional evidence to back it up. If everything checks out, it may require the involvement of the appropriate expert group to decide whether or not the record should be accepted. If it is concluded that the record is genuine, then the authority that issued the rule may need to be informed so that the rule can be adjusted to take account of the new information.

### 1.3. Summary

- Validation rules are predefined, built into the application and are always applied.
- Validation failures usually signal errors in the data that need to be corrected. The nature of the problem is usually obvious: for example a misspelled species name or an incorrectly formatted date.
- Verification rules are supplied as external rule files created by expert groups. They are obtained and kept up-to-date via the internet.
- The user chooses which verification tests they wish to apply.
- Verification failures are not necessarily errors. They should be regarded as warnings that records need further investigation. The investigation may lead to a mistake being found and corrected in the original data (like a validation failure) or to rejection of the record. If investigation shows that the record is actually correct, then the verification rule may need to be updated to take account of the new information.

## 2. Installation and updates

### 2.1. Requirements

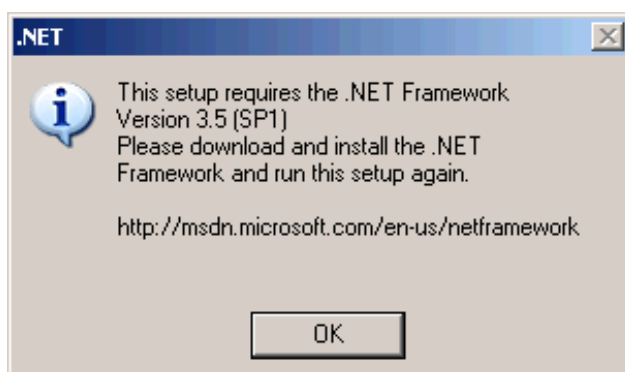
Record Cleaner is a Windows application. It is known to work under the Windows XP, Windows Vista and Windows 7 operating systems.

You need to have administration rights on the computer you are using to perform an installation. If it is your own computer at home this is not normally an issue, but if you are on an organisation's network you will probably need help from your IT administrator.

If you are using Windows Vista or Windows 7, you should run the installation file by right-clicking the installer's .exe file and then selecting "Run as administrator" from the drop-down menu (and entering the administration password if prompted). This should ensure that you have the necessary permissions to perform the installation.

#### 2.1.1. .NET framework 3.5 (SP1)

The installer will check whether or not this is available. If not, you will see a message:



You can download the installer from <http://msdn.microsoft.com/en-us/netframework/default.aspx>. Download the file dotnetfx35setup.exe (2.89Mb) and run it. You will need an active internet connection at the time you run this installer (it downloads 53Mb).

Again, on Windows Vista or Windows 7, you need to right-click on dotnetfx35setup.exe and choose to "Run as administrator" from the drop-down menu.

### 2.2. Record Cleaner installation

Run the installer and follow the on-screen prompts. The only option you can change is the location of the installation (default: C:\Program Files\Record Cleaner).

### 2.3. Software updates

You need an active internet connection for this! When updates to the software or core data files are available, you will see a link "[Software updates available](#)" on the initial screen. Click this link to automatically download and apply updates. The files which are updated by this mechanism include the software itself and its core data files, such as the master species dictionary. Please take note of any changes listed in the updates log that will be displayed when an update has completed.

If no updates are available, a message "No software updates available" (not a link) is shown.

## 2.4. Obtaining verification rules and keeping them up to date

Verification rules are created and maintained by expert groups, such as Recording Schemes and Societies, and made available via the internet. The ones you choose to use are downloaded to your computer for performance reasons. When you start the application (providing it can connect to the internet!) it will check whether updates and/or new rule-sets are available. You then decide what you want to download and install (see 8 for details).

You can also design your own verification rules by writing suitable rule files and placing them in the \VerificationData\Personal folder. Details about writing such rules are beyond the scope of this user guide: a separate guide for rule suppliers is available.

## 2.5. Limitations

The **Record Cleaner** was designed to handle dataset of up to 500,000 rows and has been successfully tested with somewhat more than this (on a machine with 3Gb of memory). However, there is a limit to the size of dataset it can process. The most likely symptom that you have exceeded this limit is that it will fail with an “Out of memory” error when loading rule files, or plotting the map, as the verification screen loads (see 7). This is most likely to be an issue if you attempt to use it to check a large *Recorder 6* (section **Error! Reference source not found.**) or MapMate database (section 10) containing more than the specified half million records.

## 3. Basic use of the Record Cleaner

**It is important to understand that the Record Cleaner DOES NOT make changes to the source data that it checks** – it only provides a mechanism to help you to locate problems which you should fix using whatever software you normally use to manage your records.

Load your data source and match up the columns it contains to particular types of data (date, grid reference, species, etc.) (see 4). You can save the results of this step in a template for re-use (see 4.5),

1. Validate your data (make sure your dates and coordinates are valid, species names recognised, etc.) (see 5),
2. Choose verification tests and apply them to your data (e.g. are the dates given for a particular species within the expected season?) (see 7),
3. Save the results of validation and verification tests to a file (see 7.5).

If errors were reported, you need to investigate them and, if corrections or rejection of whole records are found to be necessary, make these changes to the original data source. Then repeat steps 1-3 until everything is correct.

In addition:

- you must manage the downloading and updating of the verification rule files you wish to use (see 8).
- you can optionally assign each record to a vice county based upon its coordinate if you didn't specify a vice county field amongst the data you imported (see 6).
- you may use Record Cleaner to check whether your dataset is in the NBN Exchange Format prior to supplying the dataset to the NBN Gateway

## 4. Preparing and loading data

Record Cleaner can check records stored in a variety of file formats. However, the **data must be organized so that each record is stored in one line** (line of a text file, row of a spreadsheet, record of a database) **with the fields** (species name, date, etc.) **organised as columns**.

### 4.1.1. Example

Here are some records in an Excel spreadsheet:

	A	B	C	D	E	F
1	Species	Date	Location	Grid	VC	Recorder
2	Anasimymia lineata	19 May 2000	Preston Montford FSC	SJ435145	40	Ball, Stuart
3	Cheilosia variabilis	19 May 2000	Preston Montford FSC	SJ435145	40	Ball, Stuart
4	Cheilosia variabilis	21 May 2000	Earls Hill	SJ4004	40	Ball, Stuart
5	Anasimymia lineata	25 May 2000	Preston Montford FSC	SJ433144	40	Ball, Stuart
6	Cheilosia variabilis	25 May 2000	Preston Montford FSC	SJ435145	40	Ball, Stuart
7	Cheilosia variabilis	28 May 2000	Cleave	SX4068	2	Ball, Stuart
8	Anasimymia contracta	29 May 2000	Goss Moor NNR	SW9460	2	Ball, Stuart
9	Cheilosia variabilis	29 May 2000	Goss Moor NNR	SW9460	2	Ball, Stuart
10	Baccha elongata	31 May 2000	Lydyford George	SX501835	3	Ball, Stuart
11	Cheilosia variabilis	31 May 2000	Lower Creason Meadow	SX605897	3	Ball, Stuart
12	Anasimymia lineata	18 May 2002	Preston Montford FSC	SJ435145	40	Ball, Stuart
13	Anasimymia transfuga	18 May 2002	Preston Montford FSC	SJ435145	40	Ball, Stuart
14	Cheilosia variabilis	18 May 2002	Preston Montford FSC	SJ435145	40	Ball, Stuart
15	Baccha elongata	19 May 2002	Loamhole Dingle	SJ60	40	Ball, Stuart
16	Cheilosia variabilis	19 May 2002	Loamhole Dingle	SJ665055	40	Ball, Stuart

And here is the same data in a comma delimited text format (CSV file):

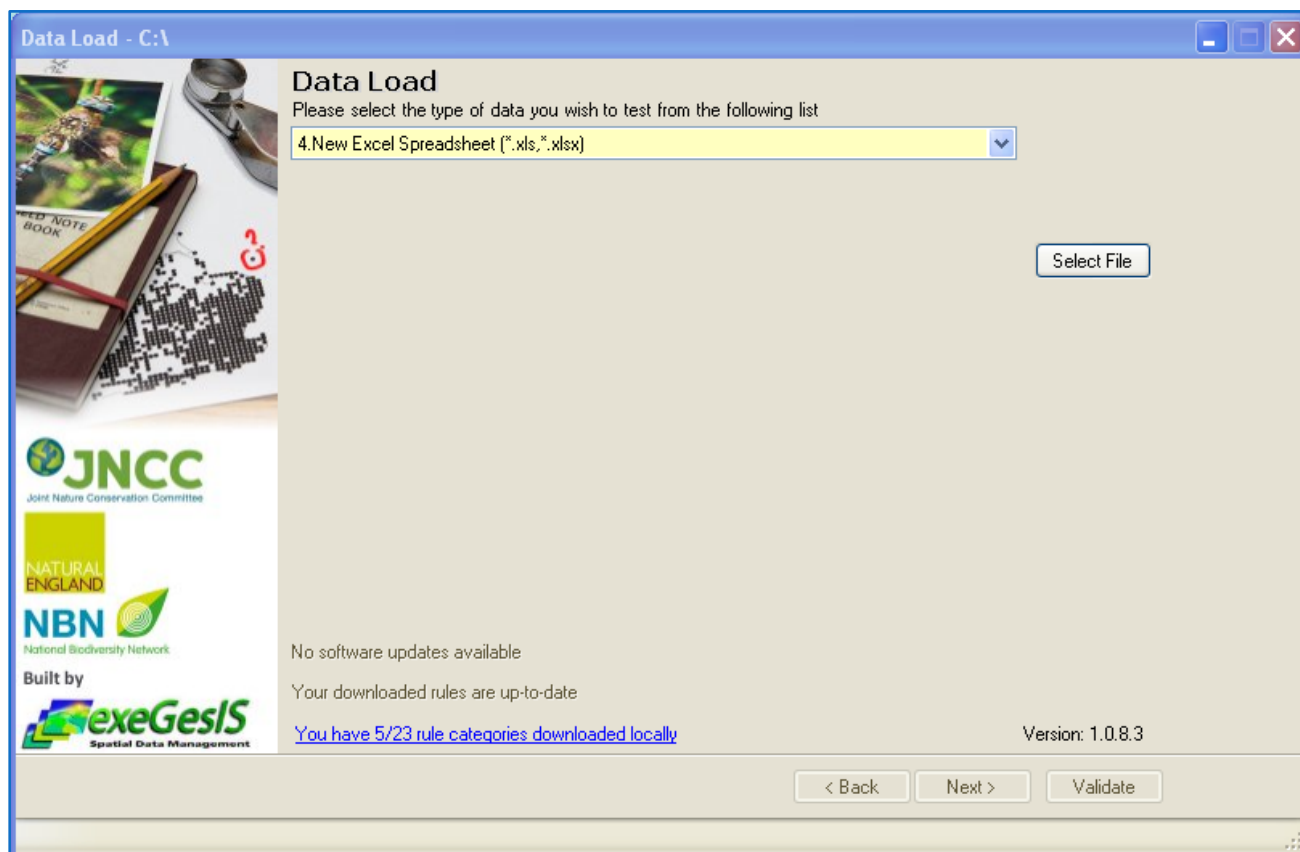
```
Species,Date,Location,Grid,VC,Recorder
Anasimymia lineata,19 May 2000,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Cheilosia variabilis,19 May 2000,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Cheilosia variabilis,21 May 2000,Earls Hill,SJ4004,40,"Ball, Stuart"
Anasimymia lineata,25 May 2000,Preston Montford FSC,SJ433144,40,"Ball, Stuart"
Cheilosia variabilis,25 May 2000,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Cheilosia variabilis,28 May 2000,Cleave,SX4068,2,"Ball, Stuart"
Anasimymia contracta,29 May 2000,Goss Moor NNR,SW9460,2,"Ball, Stuart"
Cheilosia variabilis,29 May 2000,Goss Moor NNR,SW9460,2,"Ball, Stuart"
Baccha elongata,31 May 2000,Lydyford George,SX501835,3,"Ball, Stuart"
Cheilosia variabilis,31 May 2000,Lower Creason Meadow,SX605897,3,"Ball, Stuart"
Anasimymia lineata,18 May 2002,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Anasimymia transfuga,18 May 2002,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Cheilosia variabilis,18 May 2002,Preston Montford FSC,SJ435145,40,"Ball, Stuart"
Baccha elongata,19 May 2002,Loamhole Dingle,SJ60,40,"Ball, Stuart"
Cheilosia variabilis,19 May 2002,Loamhole Dingle,SJ665055,40,"Ball, Stuart"
Baccha elongata,28 June 2003,Old Sulehay Forest,TL06119854,32,"Ball, Stuart"
```

### 4.1.2. Points to note:

1. It is a **requirement that the first row of an Excel spreadsheet or of the various forms of text file must contain names for the columns**.
2. Values in some fields in a text file may themselves contain the delimiter character (e.g. "Ball, Stuart" contains a comma – the delimiter for a CSV file). This potentially causes confusion over where the fields in that row start and end. Therefore, if you are using one of the text file formats, any field whose text contains the delimiter character must be enclosed in quotes. It does no harm to routinely enclose text fields in quotes if they contain potential delimiter characters such as commas or semicolons.

3. The data can include fields which are not normally needed by Record Cleaner (e.g. Location, Recorder), but which may be convenient to keep alongside the other data. For example, they can be included in result files and may be of assistance in tracking down errors (e.g. Having a location name visible can be very helpful when investigating a grid reference problem).

## 4.2. Loading your data



Choose the type of file you want to import from the Data Load drop-down list. The following options are initially available:

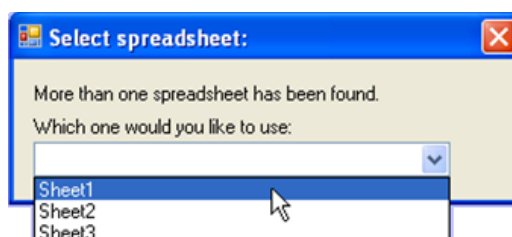
1. **New CSV File** – comma separated text
2. **New TXT File** – text file with a user defined separator character
3. **New TAB File** – tab separated text
4. **New Excel Spreadsheet (\*.xls, \*.xlsx)** – Excel spreadsheet
5. **New SQL Connection** – accesses a database using a connection string and SQL (see 9)
6. **MapMate** – reads data from MapMate's user.mdb (see 10)
7. **Recorder 6** – reads data from Recorder's NBNDData\_Data.MDF (see 11)
8. **NBN Exchange format** – tab separated text file in NBN Exchange Format (see 12)

### 4.2.1. Loading data from a spreadsheet

When you click the [Select file] button, a dialog will open which allows you to select an Excel spreadsheet. Select the file containing the data you want to check and click the [Open] button. You can select either .xls files (Excel versions prior to Excel 2007) or .xlsx files (new format introduced by Excel 2007).

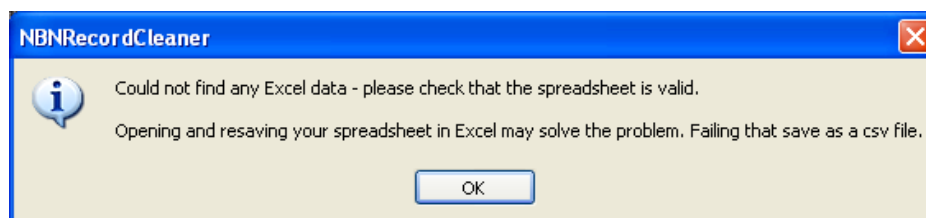


Excel files can contain a number of worksheets (by default, Excel creates new files with three empty worksheets called “Sheet 1”, “Sheet 2” and “Sheet 3”). If the spreadsheet file has more than one worksheet which is not empty, Record Cleaner needs to know which one you want to check. You will see a “Select spreadsheet” dialog containing a list of the non-empty worksheet names:



Just click on the name of one of the worksheets in the drop-down list. This step is skipped if only one worksheet has been used.

Spreadsheets that have been created and saved by a version of Excel from Excel 97 onwards should be fine, but Excel formatted files created by other software do sometimes cause problems. In this case, you may get a message:



The fix suggested in the error message, i.e. opening the file using Excel and then saving it again, usually seems to fix such files. Another, slightly more drastic, strategy is:

- Select just the rows and columns that you want to check,
- Copy this block to the clipboard (Ctrl-C),
- Create a fresh spreadsheet (File – New),
- Paste the block of data in to this new, blank spreadsheet (Ctrl-V),
- Save the new spreadsheet as a different file (File – Save as) ,
- Try loading the new file into Record Cleaner.

As a final resort, if you can open the offending file in Excel you can always save it in one of the text file formats (e.g. “Save as” - CSV).

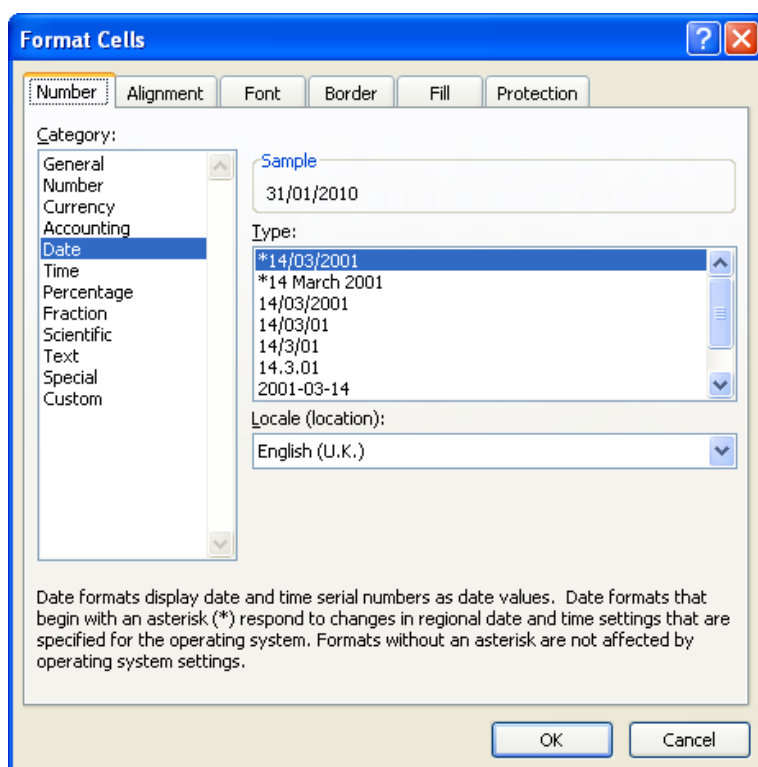
#### **4.2.1.1. Spreadsheets and dates**

Excel (and other spreadsheets and databases) store dates as numbers. Windows software stores dates and times as a real number where zero represents 01/01/1900 00:00:00 (known as the “epoch” or reference date). The integer part of the number represents a count of the number of days since the epoch (so 31/01/2010 = 40209) and the decimal part represents the time as a fraction of a day (so 31/01/2010 12:00 = 40209.5). Excel does not recognise dates before the epoch.

You normally see dates displayed in a spreadsheet in a human readable form e.g. “31/01/2010”. This is controlled by a date format applied to the cell and you can easily change it. For example:

- Type “31/01/2010” in a cell in an Excel spreadsheet,

- Right-click on the cell and select Format cells from the menu that pops up and you will see the



Format Cells dialog:

- Choose a different date format from the “Type:” list on the right (e.g. “14 March 2001” – you will need to scroll down a bit to find this entry) and click [OK]
- The cell will now show the date as “31 January 2010”

Just to verify this explanation, if you format the cell and choose the “General” item from the “Category” list on the left of this dialog, you will see the date in its raw, numeric form. The cell will show “40209” – the numbers of days since the epoch.

Excel is pretty clever about working out what you mean when you type dates into cells and applying the correct format automatically so that you see the date displayed in the way you expect. Try typing in the following into successive cells:

- 21/3/2010
- Mar 2010
- 2010
- 21/3/1810

What I see after I type in these values is as follows (you may see something slightly different depending on how you default date format is set up):

	A
1	21/03/2010
2	Mar-10
3	2010
4	21/3/1810
5	
6	

If we select these cells and change their format to “General”, we can see how Excel has interpreted them behind the scenes (and therefor what Record Cleaner will see!).

	A
1	40258
2	40238
3	2010
4	21/3/1810
5	

A couple of points to note immediately:

- A3 still contains the number 2010. Excel has not recognised this as a date and has simply stored the number as typed,
- A4 still contains “21/3/1810”. Excel hasn’t reconised this as a date (because it does not recognise dates before the epoch) and, since it isn’t a number either, it has stored it as text. The clue is that it is left-justified in the cell (compare with the numbers in A1 – A3 which are right-justified).

Now selecting these cells again and turning date formatting back on again, we see:

	A	B
1	21/03/2010	
2	01/03/2010	
3	02/07/1905	
4	21/3/1810	
5		
6		

Spot the problems!

- A2 now contains “01/03/2010” instead of “Mar-2010”. If you enter a date as “<month> <year>”, Excel will format it in “mmm-yyyy” format, but behind the scenes, it actually stores the number representing “01 <month> <year>”.
- A3 now contains a date – “02/07/1905”. This is the date represented by 2010 days after 01/01/1900!

And this is how Record Cleaner will see it! It will quite happily notice that “21/3/1810” is text and handle it correctly, but because “Mar 2010” and “2010” are numbers in a date column, it will interpret them as day numbers and understand “01/03/2010” and “02/07/1905”.

You can make sure that Record Cleaner interprets the non-standard dates (i.e those that are not a full “mm/dd/yy” dates) correctly by ensuring that they are stored as text NOT numbers. To achieve this, just prefix what you type with an apostrophe in these cases. To see this in action, start again with an empty spreadsheet and type in successive cells:

- 21/3/2010 -- no apostrophe – this really is a date
- 'Mar 2010 -- prefix with an apostrophe to tell Excel to treat it as text
- '2010 -- prefix with an apostrophe to tell Excel to treat it as text
- 21/3/1810 -- no apostrophe – this really is a date

And go through the same exercise of switching them to general format and back to date format:

	A
1	21/03/2010
2	Mar 2010
3	2010
4	21/3/1810
5	

after input

	A	B
1	40258	
2	Mar 2010	
3	2010	
4	21/3/1810	
5		

apply "General" format

	A	B
1	21/03/2010	
2	Mar 2010	
3	2010	
4	21/3/1810	
5		

apply "Date" format

Because A2 and A3 were explicitly entered as text (note that they appear left-justified in the cells), the format changes don't affect them. But Excel has noticed that you entered a number as a string in A3 – which is why it puts the green corner on the cell. If you select this cell, you will see a warning about this:

	A	B	C	D	E	F	G
1	21/03/2010						
2	Mar 2010						
3	2010						
4	21/3/1810						
5							
6							

Record Cleaner will "notice" that these cells in the date column contain text (not numbers) and convert them from text to dates correctly. The moral of the story is that, if you want to mix different types of dates in a single column in a spreadsheet, make sure that anything that is not an ordinary date is text!

Note that similar problems may occur when database software such as *Recorder* or MapMate import spreadsheet data. Like Record Cleaner, they only see the behind-the-scenes number. In the Unix operating system the epoch is 01/01/1970 00:00:00, so there can be additional issues about moving dates from Unix (including Mac OS) software to Windows.

#### 4.2.2. Loading data from text files

If you select "New TXT file" (i.e. a text file delimited by some arbitrary character) then you will see an extra prompt for the delimiting character:

**Data Load**

Please select the type of data you wish to test from the following list

2.New TXT File

Please enter the delimiting character #

Select File

Select the delimiting character from the drop-down list. You are given a choice of: TAB, hash sign (#), comma, semicolon or hyphen, or you can type some other single character in the box.

If you select “New CSV file” (i.e. comma delimited text) or “New TAB file” (i.e. tab delimited text) then the delimiter character is known and so this extra prompt is not shown.

Click the [Select file] button and you will see an open file dialog. Choose the file you want to load and click [Open].

Click [Next >] to continue.

### 4.2.3. Viewing the data file you have chosen

Once you have selected your input file, a “[View data file](#)” link will appear at the bottom, left-hand corner of each of the subsequent screens. You can click this link at any time to view the raw data from the file you have loaded. For example, if you loaded a spreadsheet, then clicking this link will open Excel with the file loaded. If you loaded a text file, then clicking the link will open Notepad with the file loaded.

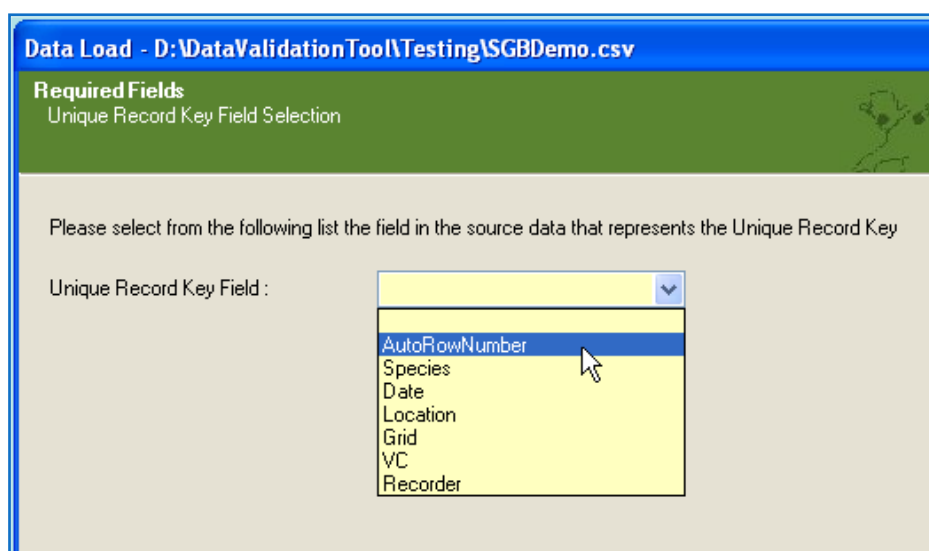
## 4.3. Matching columns

In the next few steps, you are prompted to tell Record Cleaner which columns in your file contain the various types of data it needs (date, spatial coordinates, species names, etc).

### 4.3.1. Unique record identifier

You are prompted to identify the column containing a unique record key. You need this because the order of data rows may change when you view data (e.g. if you decide to sort it by date or species name) or you may only see subsets of the rows (e.g. those with an error). So you need some way of identifying the records in the original source file.

The drop-down list shows the column names read from the first line of the input file. Although database tables often contain an unique key column (e.g. *Recorder* 's TAXON\_OCCURRENCE\_KEY), spreadsheets and text files usually do not. In these cases, you are offered the option “AutoRowNumber”. This will use the row address in the spreadsheet or the line number in the text file, as the identifier.



Click on the appropriate field name in the list and then click the [Next >] button.

### 4.3.2. Date

You are prompted to identify the date column. If you simply have a date stored in a single column in your input file, click on the column name in the drop-down list.

**Required Fields**  
Date Field Selection

Please select from the following list the field in the source data that represents the date information

Date :

- Species
- Date**
- Location
- Grid
- VC
- Recorder

Record Cleaner handles dates in the NBN vague date format. Any of the following formats are acceptable:

12 Sep 2003 12/09/2003 12-09-03	A full date in “day month year” format. Most special characters are acceptable as punctuation (but not the hyphen, “-”, which is used as the delimiter for date ranges). Month names or numbers can be used. Two digit years are interpreted as 21 <sup>st</sup> Century if the number is <40, or 20 <sup>th</sup> Century otherwise (i.e. 12/9/5 is interpreted as 2005, whilst 12/9/65 is interpreted as 1965)
Mar 2007	Month and year
2007	Year
12/9/2007 – 16/9/2007 Mar 65 – Jul 67 2003-2005	Date range. A hyphen (-) should be used to separate the two dates and should NOT be used as punctuation in dates.
-1987	An open ended range (“before or during 1987”). “1987-” (“1987 onwards”) will also be accepted, but is deprecated. Use something like 1987-2010 (i.e. close the range with the current year).

#### 4.3.2.1. More complex dates

Record Cleaner can also handle more complex situations in which date ranges are stored in multiple fields. There is a link [“Show advanced date settings”](#) which provides access to additional fields.

**Required Fields**  
Date Field Selection

Please select from the following list the field in the source data that represents the date information

Start date: StartDate

End Date : EndDate

Vague date type:

If the data source contains a field describing the vague date type then please select it

Date Type :

- RecNo
- Species
- Date
- GridRef
- VC
- Recorder
- Date Type**

#### 4.3.2.2. Day, month and year in separate columns

Record Cleaner does not have facilities to handle dates stored in this way. However, the source data can be easily modified by loading it into a spreadsheet and using the DATE(y,m,d) function to combine the separate columns into a single “date” column which Record Cleaner will be able to understand.

D2		fx =DATE(C2,B2,A2)			
	A	B	C	D	E
1	day	month	year	date	
2	23	3	2010	23/03/2010	
3	16	7	2005	16/07/2005	

Click the [Next >] button to proceed.

#### 4.3.3. Spatial coordinate

You are prompted to identify the column(s) containing the coordinates of the observation. You need to give at least two pieces of information: identify the Coordinate type and then select one or two columns from your data which contain the coordinates.

**Required Fields**  
Coordinates Fields Selection

In this section you must select either one or both fields that represent the Coordinates data. Then you must either select a field that represents the Coordinates data type or an actual Coordinates data type from the list. [Show](#)

Coordinate type (must select one of the options below)

Select the coordinate type:  
Actual Coordinates Type : British gridref (e.g. SM123456) or E/N

Coordinates

Enter Gridref or Easting and Northing (if combined in one field) or Easting

Enter Northing (if previous field is Easting and Northing)

Species  
Location  
Grid  
VC  
Recorder

[View data file](#) [< Back](#) [Next >](#)

##### 4.3.3.1. Actual coordinate type

Select the type of spatial coordinate your data contains. The options are:

Option	Use for
British gridref (e.g. SM123456) or E/N	OSGB grid references as used on Ordnance Survey maps either in the usual alphanumeric format (SM123456) or as easting and northing coordinates in km.
Irish gridref (e.g. S123456) or E/N	OSNI grid references as used on Ordnance Survey Ireland maps either in the usual alphanumeric format (S123456) or as easting and northing coordinates in km.

Channel Island Grid (e.g. WA/WV)	Grid references from the Channel Islands that use “WA” or “WV” as their initial letters either in the usual alphanumeric format (WA123456) or as easting and northing coordinates in km.
British, Irish or CI gridref (Not E/N)	This option allows for data containing mix of grid references from the three grid reference systems in use in the British Isles and will attempt to identify which is appropriate from the data. (i.e. if the grid reference starts with only one letter then it is Irish, if it starts with two letters then, if they are WA or WV then it is from the Channel Island, otherwise it is from GB). You cannot use easting, northing coordinates in this case because they do not provide any clue to which of the grid reference systems they belong.
Lat/Long (WGS84)	Latitude and longitude based on the WGS84 global ellipsoid. Lat/Long coordinates obtained from GPS recorders are usually in this format.
British Lat/Long (OSGB36)	Latitude and longitude as used by the Ordnance Survey for Great Britain. Lat/Long read from OS maps are in this format.
Irish Lat/Long (IRENET75)	Latitude and longitude as used by the Ordnance Survey Ireland. Lat/Long read from OS Ireland maps are in this format.
UTM Zone 30N	UTM coordinates (in metres) within Zone 30N (which covers Britain).

#### 4.3.3.2. Gridref or Eastings and Northings

If your coordinate type is a grid reference you only need to select a column name in the first of the two coordinate fields. This will be your column containing the grid reference as a string (e.g. “SM123456”). The second coordinate field “Enter Northing ...” should be left blank in this case. However, you can optionally provide grid references in two fields containing the easting (x) and northing (y) coordinates in kilometres. In that case you would need to select a column name in both of the coordinate fields.

Other points to note about grid references:

- Tetrads using the “DINTY” system are accepted (e.g. TL29S, B36G)
- Pentads (5km squares) are accepted (e.g. TL29SW, TL29NW, TL29NE, TL29NE)
- Spaces in grid references are ignored (e.g. “TL 29 88”, “TL 2988” and “TL2988” are considered equivalent).

If the coordinate type is one of the Lat/Long options (e.g. “Lat/Long (WGS84)”) then two columns (containing the latitude and longitude coordinates) are always required.

#### 4.3.3.3. All numeric grid references

BRC used to advocate grid references in an all numeric format : e.g. 52/203998 instead of TL203998 (i.e. the 100km grid square is given as a number, rather than by two letters, and separated from the coordinates by a slash). This format is no longer recommended and is not supported by Record Cleaner.



#### 4.3.3.4. More complex coordinates

Record Cleaner can also handle more complex situations in which additional information - like the precision of the coordinate - is available. There is a link "[Show advanced coordinate settings](#)" which provides access to additional settings.

**Required Fields**  
Coordinates Fields Selection

In this section you must select either one or both fields that represent the Coordinates data.  
Then you must either select a field that represents the Coordinates data type or an actual Coordinates data type from the list. [Hide advanced coordinate settings](#)

**Coordinate type (must select one of the options below)**  
Select the coordinate type:  
Actual Coordinates Type : Lat/Long (WGS84)

Or if the data contains a field that defines the coordinate type  
select it from the list below.  
Coordinates Type Field :

**Coordinate precision (must select one of the options below)**  
If the data contains a precision field then you can select it here  
Coordinates Precision Field : Precision

If your data contains GridRef the system can determine the precision from these  
Precision Auto Detect : ☐ ONLY for OSGB , Irish or CI grid references

Alternatively you may set the precision of the coordinates manually  
Manual Precision setting (Metres): 100

**Coordinates**  
Enter Longitude or Longitude and Latitude (if combined in one field)  
Lat

Enter Latitude (if previous field contained  
Long

In this example, the spatial coordinate data (from an Excel spreadsheet) is organised like this:

E	F	G
LAT	LONG	Precision
54.1895801	-2.793937927	100
54.18866111	-2.796985708	100
54.1706018	-3.233275128	100
54.18594453	-2.799998649	100
54.17801601	-2.775328795	100
52.47633236	0.677205804	100
52.19008862	0.951189404	100
51.41946442	-2.115061017	1000
50.98856615	0.009201262	1000

LAT and LONG are in decimal degrees and the Precision in metres. The "Actual Coordinates Type" has been declared as Lat/Long using the WGS84 datum (the usual system for GPS receivers) and "Precision" has been declared in the Coordinate Precision Field. Unlike grid references (which have an implicit precision determined by the number of figures that are quoted), Lat/Long coordinates indicate a point with no implied precision. It is therefore necessary to state the precision explicitly either as a field in the data as here – (e.g. the measurement accuracy recorded by a GPS receiver) or to give a single figure which is assumed to be the same for all rows (this single value would be entered in the "Manual Precision setting (Metres)" field - which defaults to 100).

Click the [Next >] button to proceed.

#### 4.3.4. Species

You are prompted to identify the column containing the species name.

The screenshot shows a dialog box titled 'Required Fields' with a subtitle 'Species Field Selection'. The background is green with a map of the UK. The main area is light beige and contains the text: 'In this section you may select either a Species field or a taxon version key field'. Below this, it says 'Select the species from either:'. There are two labels: 'Species Name:' and 'Taxon Version Key:'. The 'Species Name:' label is next to a yellow dropdown menu. The 'Taxon Version Key:' label is next to a yellow dropdown menu that is open, showing a list of options: 'Species', 'Location', 'VC', and 'Recorder'.

Most of the time, you will select the name of the column which contains the scientific name of the species from the Species Name drop-down box. Record Cleaner will attempt to look up TAXON\_VERSION\_KEYS corresponding to the names from the master species list (which is derived from the NBN taxon dictionary).

If you already have a TAXON\_VERSION\_KEY in your data, you can specify this instead by choosing the column name from the Taxon Version Key drop-down box. In this case, Record Cleaner will simply check that the key exists in the master species list.

Click the [Next >] button to proceed.

#### 4.3.5. Vice-county

Finally you are prompted to identify a column containing vice-county numbers or names.

The screenshot shows a dialog box titled 'Optional field' with a subtitle 'Vice County field'. The background is green. The main area is light beige and contains the text: 'This field is optional. If the data contains a Vice County field then you may select this from the list below. If you do select a field then it must contain valid names or numbers.' Below this, it says 'Vice County:'. There is a label 'Vice County:' next to a white dropdown menu. The dropdown menu is open, showing a list of options: 'Location', 'VC', and 'Recorder'. The 'VC' option is highlighted in blue. A mouse cursor is pointing at the 'VC' option.

This field is optional but highly recommended, because the grid reference is one of the most error prone field in biological records! Two common mistakes (often made by even the most experienced recorder!) are: giving the wrong 100km square letters or reversing the easting and northing when looking up a grid reference from the map. The first of these even occurs not uncommonly when transcribing grid references from a GPS! If you give both a grid reference and a vice-county, then it is possible for Record Cleaner to test whether the grid reference lies within the vice-county boundary. This will nearly always pick up these two common errors.

Click the [Next >] button to proceed.

#### 4.4. Including additional fields

The “Optional fields” screen allows you to include additional columns alongside your other data. These will not be used in the validation process, but are kept with the rest of the data, are displayed on the reporting screens and can be included in output files. This is often useful. For example, if an error is reported by the “coordinate in Vice-county” check, a location name is very helpful in determining whether it is the grid reference or the VC number that is incorrect.

Optional fields can be used by verification rules. For example, if your dataset is supposed to include a DAFOR abundance code, it is possible to provide a rule file to check that the code is one of the recognised values. Similarly, if a dataset contained information on life stage (e.g. “larva”, “adult”), this could be taken into account by a rule that checked that the date was within an expected season of the year (i.e. different acceptable start and end dates apply depending on the value in the life stage column). In these cases, for the verification to occur as expected, the author of the rule file would need to specify the name of the data columns from which their rule expected to get this additional information (e.g. “Abundance”, “Stage”) and you would need to ensure that you labelled the columns correctly in your data file.

To include an additional field in the dataset, click on its name in the left-hand list so that it is highlighted and then click the [Add ->] button – or, more simply, just double-click the name. The name will be transferred from the left-hand to the right-hand list and will appear in the grey-box on the right which is a list of the fields included in the internal, working dataset.

**Optional Fields**  
User Selected Fields for inclusion in the dataset

This section allows you to select additional fields, other than mandatory fields, so that you may include these additional fields in the dataset. Fields are added by selecting them from the list on the left and pressing the 'add' button, or by double clicking on the desired field. They can be removed by selecting them from the list on the Top right and pressing the 'Remove' button. The list on the bottom right is a list of the fields already included in the dataset

Location

Add ->

<- Remove

Recorder

Click [Next >] when you are ready.

#### 4.5. Saving a template

You now get a chance to save all the setting you have made so far.

You are now ready to validate your data file.

Before you press the validate button you may want to save your template for future use.

Enter the name of you template and press the 'Save Template' button.

If there is already a template with the same name you will be asked if you want to overwrite.

Once saved the template will be available to re-use from the list of imports at the beginning of this wizard.

Template Name :

Enter a name for the template and then click the [Save template] button. Note that, although the field is pre-filled with the name of the data type you chose back in the first screen, you cannot save it using this name. You need to provide a new name in order to save it.

#### 4.5.1. Using templates

Once you have saved a template, the name you gave to it will appear in the list of data types in the first screen of the application. Choosing a template will load all the settings you saved and give you the option to skip the column matching steps. This facility is useful in the following cases:

- If some of the records don't validate. You may want to edit the original data file. Once the problems have been fixed, you want to re-run the checks. If you saved a template, then you just need to select it in the initial screen and you can run the validation tests straight away without needing to go through the column matching steps again.
- If a regular contributor always submits records in the same format, then you can set up a template for use whenever you receive another batch.

This screenshot shows the initial screen with a template selected. The [Validate] button is now active. This allows you to run the validation tests immediately, by passing the column matching screens. At this point, you can also use the [Select File] option to select a different file to check. This allows the use of a template to check any number of different source files that are stored in the same format.

Data Load - C:\NBN Record Cleaner demonstration dataset.xls

**Data Load**

Please select the type of data you wish to test from the following list

You may now go directly to the data validation or continue through the wizard to make changes

You may change the file you are validating so long as the structure is identical

**JNCC**  
Joint Nature Conservation Committee

**NATURAL ENGLAND**

**NBN**  
National Biodiversity Network

Built by  
**exeGesIS**  
Spatial Data Management

[View data file](#)

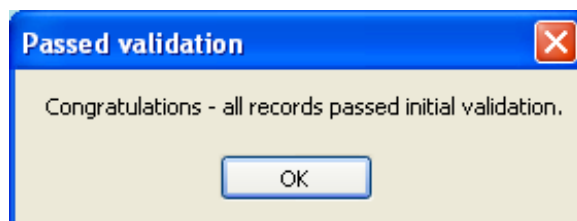
No software updates available  
Your downloaded rules are up-to-date  
[You have 5/23 rule categories downloaded locally](#)

Version: 1.0.8.3

## 5. Validation

A series of progress messages are shown as validation takes place. This is a fairly quick process and, for a small dataset, should take only a few seconds. Even for large datasets (up to 500,000 rows have been tested) it should only take a minute or so.

Assuming all the tests pass, you will see this message and you can click [OK] to proceed.



However, if validation errors were found you will see the Data Cleansing Form:

A screenshot of the 'Data Cleansing Form' window. The window has a blue title bar and a menu bar. Below the menu bar is a toolbar with buttons: 'Failed records', 'Filter', 'Apply filter', 'Advanced Filters', 'Reload Original Data', 'Re-run Validation', 'Save failed records', 'Export edited records', and 'Run Verification'. The main area is a table with columns: 'ErrorDesc', 'AutoRowNumber', 'Gridref', 'Taxon', 'VC', 'Site', and 'TVKFIELD'. The table is titled 'Unknown Coordinate Type' in red. Several rows are highlighted in red, indicating errors. Annotations with yellow boxes point to various parts of the interface: 'Choose between viewing all records or only those with errors' points to the 'Failed records' dropdown; 'Summary showing no. of errors' points to the '10 of 5293 records' text; 'Description of the error' points to the 'ErrorDesc' column; 'If there are a lot of errors, you can filter them to examine a few at once - e.g. just those relating to dates' points to the 'Filter' and 'Apply filter' buttons; and 'The row identifier - allows you to refer back to the original data source' points to the 'AutoRowNumber' column.

Choose between viewing all records or only those with errors

Summary showing no. of errors

Failed records ▾ Filter :  for  Apply filter Advanced Filters 10 of 5293 records

Reload Original Data Re-run Validation Save failed records Export edited records Run Verification

**Unknown Coordinate Type**

ErrorDesc	AutoRowNumber	Gridref	Taxon	VC	Site	TVKFIELD
Unknown Scientific Name	5175	SE1447	Vicia craca	64	Ben Rhydding Gravel Pits	
Unknown Scientific Name	5177	SE1656	Oxalis acetosela	64	Washburn	
Unknown Coordinate Type	5174		Sedum acre	64	Starbotton	NBNSYS00
Unknown Coordinate Type	5176	SD09968				NBNSYS00
Unknown Coordinate Type	5181	SSD935793				NHMSYS00
Start date is in the future	5183	SE0755				NHMSYS00
Not a valid vague date	5173	SD955700				NBNSYS00
Not a valid vague date	5180	SE0453	Potamogeton polygonif...	64	Bolton Abbey Railway Stati...	NHMSYS00
Non Unique Scientific Name	5178	SE1853	Redshank	64	Washburn	
	5179	SE2248	Redshank	64	Washburn	

Description of the error

If there are a lot of errors, you can filter them to examine a few at once - e.g. just those relating to dates

The row identifier - allows you to refer back to the original data source

This screen includes a variety of tools to help you locate and fix validation errors. When it opens, it will only show the failed records (in red). The number of records which failed is shown at the top-right (10 out of 5,293 in this case). The drop-down button (entitled "Failed records") at the top-left corner has options to show all records and the "Filter" controls allow you to further refine which rows are shown. If you choose to view all records, the rows without errors are shown in black and have nothing in the "ErrorDesc" column. The "ErrorDesc" column shows a description of the error found in each row.

## 5.1. Fixing errors

You can fix errors by editing the data in the Data Cleansing Form, but this **WILL NOT** correct the **original file** from which the data came originally.

Unknown Coordinate Type									
ErrorDesc	AutoRowNumber	GridRef	VC	TVKFIELD	Species	SPECIESID	_Precision	DateFrom	
Non Unique Scientific Name	2	ND268893	111		Platycheirus scut...		100	26/06/1982	
Not a valid vague date	6	NS807966	86	NBNSYS000000...	Chrysotoxum arc...	1	100		
Unknown Scientific Name	8	TL193985	31		Bacha elongata		100	15/05/1994	
Unknown Coordinate Type	24	TF120	32	NBNSYS000000...	Leucozona lucor...	1		04/06/2005	

In the above example, we have four rows with errors reported. The bottom one (highlighted) has an “Unknown coordinate type” error because the grid reference has only three figures (TF120). Some investigation shows that the last digit is missing and it should be “TF1201”. We can click on the cell containing the incorrect grid reference to select it and then edit its content to make the correction.

The first row has the error “Non Unique Scientific Name” for the species “Platycheirus scutatus”. This is because this species was split and there are now two different meanings of this name (an aggregate before the split and a segregate afterwards). Right-clicking on this name allows us to choose between the two meanings. A drop-down list is shown containing the authorities and other information about the two alternatives which were found in the species dictionary for this name:

Species	VC	HasError	PTVKFIELD	TVKFIELD	S
Platycheirus scut...	111	<input checked="" type="checkbox"/>			
Chrysotoxum arc...		<input type="checkbox"/>	<input checked="" type="checkbox"/>	NBNSYS000000...	1
Bacha elongata		<input type="checkbox"/>			
Leucozona lucor...	32	<input checked="" type="checkbox"/>	NBNSYS000000...	NBNSYS000000...	1

Choose one of the options from the drop-down list and click ☒ to select the correct meaning of the name.

We also have the error “Unknown Scientific Name” in the third row. This is because the name is misspelt. It should be “Baccha elongata”. Again, we can right click on the cell and we get a popup window which allows us to lookup the correct name. Type “Baccha” into the box, drop-down the list of matches that it finds, and select the correct one from the list.

15-May-94	TL193985	Bacha elongata	31	<input checked="" type="checkbox"/>
04-Jun-05	TF1201	Leucozona lucor...	32	<input checked="" type="checkbox"/>

List names starting with: BAC

Baccha

Baccha elongata

Baccha obscuripennis

Bacidia absistens

Bacidia amoldiana forma corticola

Bacidia assulata

Bacidia auerswaldii

Bacidia bagliettoana

Again we finish the update by clicking  to return the correct name to the grid.

We have now fixed three errors. Click the [Re-run Validation] button in the toolbar and our corrected records will be re-validated.

Original invalid dates may not be displayed in the data cleansing form so that they can not be corrected directly in the form. Invalid dates may however be saved along with other failed rows so that they can be corrected in the original dataset.

## 5.2. Saving failed rows and rows you have edited

The [Save failed records] button allows you to save a tab delimited text file containing just the rows from the original data file which failed validation. An extra field is appended to each row containing the error description message. Clicking this button shows a message:



When you click [OK] you will see a file save dialog which allows you to name the file and choose the directory to which it will be written.

Here is an example of the output (loaded into Notepad):

AutoRowNumber	Species	Date	Location	Grid	VC	Recorder	ValidationErrors
10	Bacha elongata	31/05/2000	Lydyford George	SX501835	3	Ball, Stuart	Unknown Scientific Name
11	Cheilosia variabilis	31/05/2000	Lower Creason Meadow	SX605893	3	Ball, Stuart	Unknown Coordinate Type
18	Brachypalpoides tentus	01/01/0001	Old Sulehay Forest	TL06119854	32	Ball, Stuart	Not a valid vague date

The [Export edited records] button is relevant if you have edited rows in the grid to correct errors. It behaves in a similar way to the [Save failed records] button and saves a similar tab delimited text file, but the information that is saved consists of the row number and column name of the item you edited and the new value that you entered.

Here is an example of the output (loaded into Notepad):

"AutoRowNumber"	"FieldName"	"NewValue"
"11"	"Grid"	"SX605893"
"18"	"Date"	"28 JUNE 2003"

This is intended as a record of your actions so that you could apply the same fixes to the original data file. These files are particularly useful if you received the file from somebody else. Either file could be emailed back to the original author so that errors can be fixed at source.

## 5.3. Fixing the original data file

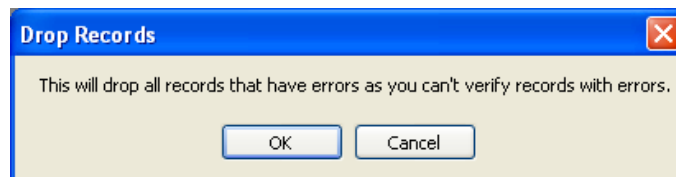
It is recommended that the original data file is corrected and reloaded rather than editing the data in the grid in the Data Cleansing Form. To do this, open the original file using the appropriate software (e.g. Excel for a spreadsheet, Notepad for a text file) and make corrections there. The row numbers in the original file are shown in the "AutoRowNumber" column to make it easier to find the offending items. Once the corrections are done and the edited version of the original data file has been saved, click the [Reload Original Data] button in the toolbar to reload and re-validate the data.

**Note that, if you reload your original data, any edits to dates, spatial references, etc. that you have made in the grid in the Data Cleansing form will be over-written and lost.**

The exception to this is species name matching. For example, if you had an “Unknown scientific name” error, and have looked up the correct name using the facilities in the grid, that “match” is stored as part of your user configuration and will be re-applied if the file is reloaded. Therefore such names will remain corrected even if you reload the original data.

#### 5.4. Proceed without fixing errors

You can proceed without fixing the errors - although any rows that still fail validation will be dropped from the internal, working dataset at this point. To proceed, click the [Run Verification] button in the toolbar. You will see a warning:



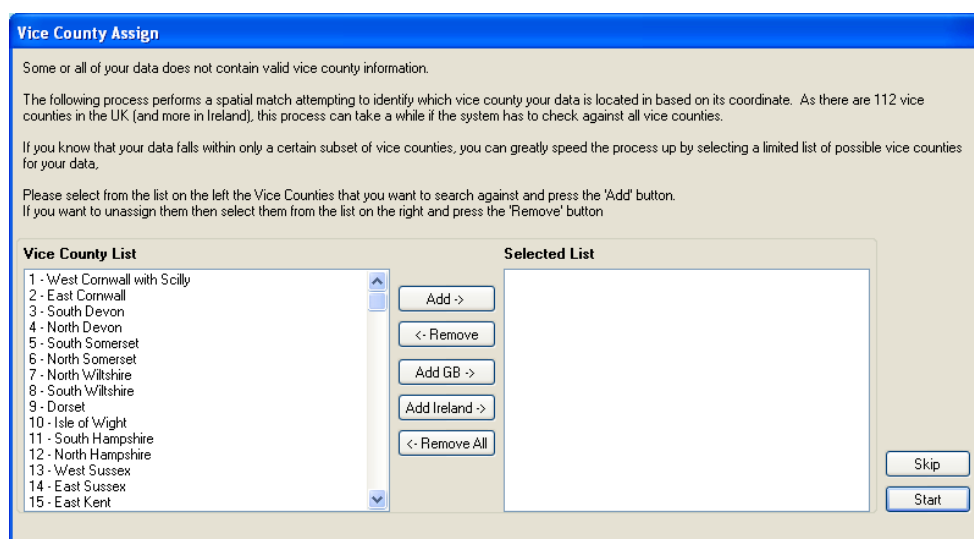
If you click [Yes] to continue you will see something like:



And then verification will be run on the remaining rows that passed the validation tests.

### 6. Allocating Vice-counties

If you did not select a Vice-county field in your data then you will see a screen that allows you to allocate vice-counties, based on the spatial coordinates of the records. This screen is displayed after validation has been completed and before the verification screen is shown.



If you don't want to allocate VCs you can simply click the [Skip] button and the verification screen will be loaded immediately.



If you would like to allocate VCs, then you need to choose the list of Vice-counties which will be checked to see whether your records fall within their boundaries. This is quite a computationally intensive task, so it will speed things up if you select as few VCs as possible to check against. Assuming you know roughly where your records are located, it is usually possible to select a subset of VCs.

To select the Vice-counties against which records will be tested:

- Double-click on names in either list. The item you double clicked will be transferred to the other list.
- Select some items by clicking on their names in one of the lists. You can use the usual Windows Shift-click and Ctrl-click methods to select multiple items.
  - To select a block of adjacent items: Click on the first item of the block. Hold down the Shift key and click on the last item of the block. All items between the two you clicked should be selected.
  - To select several items that are not adjacent: hold down the Ctrl key whilst clicking the items you want. Keep clicking whilst holding down Ctrl to select as many items as you want.
- Click the [Add ->] button to transfer selected items from the Vice County List to the Selected list, or the [<- Remove] button to transfer selected items out of the Selected List.
- Add predefined lists of all GB or All Irish VCs to the Selected List by clicking the [Add GB ->] or [Add Ireland ->] buttons.
- Clear the Selected list by clicking the [<- Remove All] button.

When the Vice-counties you want to test against are listed in the Selected List, click the [Start] button to begin the process. A progress bar will be displayed along the bottom of the screen.

When the process has finished, the Verification screen will be displayed and there will be a VC field in the Record grid showing the allocated VC number. An item **“VCFIELD” will also be present** in the Save Results tab allowing you to save the allocated VC numbers alongside your other data.

### 6.1. How VCs are allocated

The spatial coordinates in a given record are checked against a lookup list (DominantVC.txt) which lists the VCs that overlap grid squares at 10km, 2km and 1km resolution.

- The grid square in which the coordinate lies is identified. If the precision of the coordinate is 1km square or better, this will be a 1km square, otherwise a 2km or 10km square is used -depending on the precision of the spatial coordinate.
- If only one VC overlaps the grid square, then that VC is allocated.
- If the precision of the spatial coordinate is 1km or less and more than one VC overlaps the grid square, then the VC with the greatest area of overlap with the grid square is allocated (i.e. the “dominant VC”).
- If the coordinate is of greater precision than 1km and more than one VC overlaps the 1km square, then point-in-polygon tests are done against the Vice-county boundaries to find out in which of the two or more overlapping VCs the coordinate falls. (This is the most time consuming test, so it is only used when really necessary!)
- If a spatial coordinate does not fall within one of the grid squares listed in the lookup list (i.e. the grid square is not within a British or Irish VC), then no VC number will be allocated. The VC field in the Result grid will be blank. This will be the result for coordinates that are not on land.

## 7. Verification

You will then see a progress message displayed as the available rules are loaded and the map is prepared. This can take a while since there may be a large numbers of rule files to load and spatial references to map. Once this has been done, you will see the Verifications Tests screen. This screen lists the records in a grid, shows them on a map and provides tools to choose the verification tests that you want to apply.

**Rule selection tree**

**Map showing record locations**

**Toolbar to control the map**

**Detailed results for the currently selected record**

**Currently selected record**

**Grid listing records**

**Button to start verification tests**

**Tabs to switch between record grid, summary of test results and screen to save result files.**

ID	DIFF	AUTOWOWNUMBI	GRIDREF	Species	Count
2	2		HY20		
3	3		HY2001	Cheilosia bergens...	
2	4		HY2001	Melanostoma sca...	111
1	5		ND263888	Rhingia campestris	111
2	6		ND265887	Helophilus pendu...	111
2	7		ND2688	Anasimyia lineata	111
2	8		ND2688	Eristalis arbustorum	111
2	9		ND2688	Helophilus pendu...	111
3	10		ND2688	Lejogaster metalli...	111
			ND2688	Melanooaster hirt...	111

The way you work with this screen is:

- Select one or more rules you want to apply by ticking boxes in the left-hand pane,
- Apply the selected rules to the records by clicking the [Start Test] button at the bottom of the left-hand pane,
- Use the “Summary” and “Records” tabs above the grid, as well as the map, to view the results of the tests,
- Use the “Save Results” tab to save the test results to a file.

## 7.1. Identification difficulty

The “ID DIFF” column flags the difficulty of identification of the species in each row. These can be supplied (using the same mechanism as rule files) by expert groups and allow species to be classified according to identification difficulty on a scale of 1-5, where 1 is easiest and 5 is most difficult. Expert groups decide whether or not they want to use all five available levels and exactly what each level means for their group. Record Cleaner uses colour coding to give the user a quick visual indication. The colours range from green for category 1, through yellow and orange, to red for category 5. If a species has not been classified, it will be listed as “0” (meaning “unassigned”) and no colour will be shown.


## 7.2. Mapping


The map shows the spatial references of the records as boxes plotted on a map of the British Isles. If the spatial references are grid references, then these boxes show the actual grid square in each record – so 10km squares, tetrads and 1km squares are plotted as different sized boxes. The boxes are semi-transparent so you can see where they overlap. Grid references of greater precision than a 1km square are plotted as 1km squares – otherwise they would be too small to show on the map.


The toolbar above the map contains a number of tools to manipulate the map.

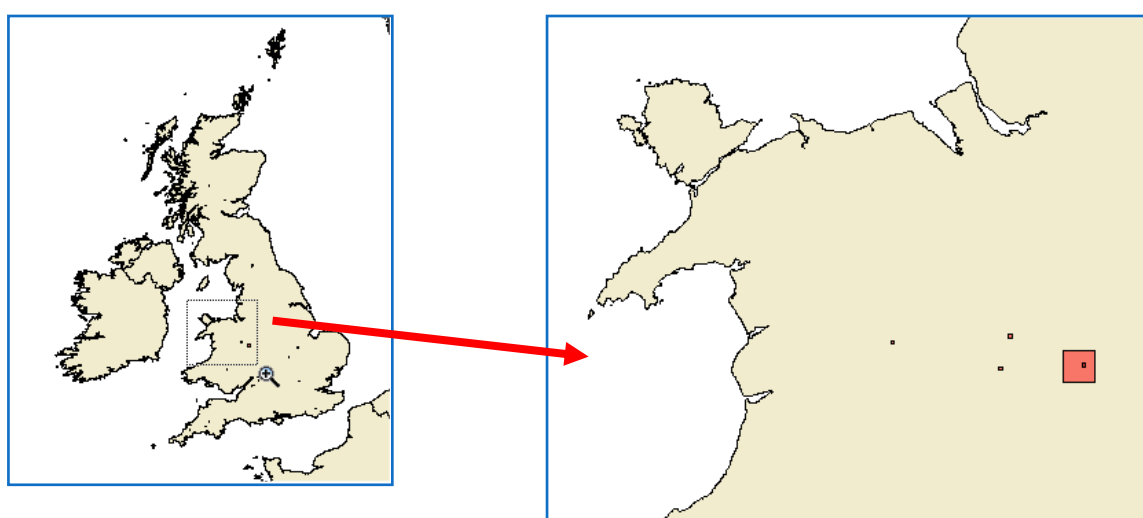
### 7.2.1. Zooming in and out, pan and zoom to extent

These tools should be pretty obvious and intuitive.

The zoom in tool  can be used to draw a rectangle over an area of the map. The map will then be redrawn so that the selected rectangle is maximized in the screen. Just clicking will double the scale centred at the point where you clicked. If you right click, or right-drag, with this tool selected, it will zoom out instead.



The zoom to extent button  will redraw the whole map as it first appeared (i.e. zoomed to a scale chosen so that the map encompasses all the spatial references present in the data set).

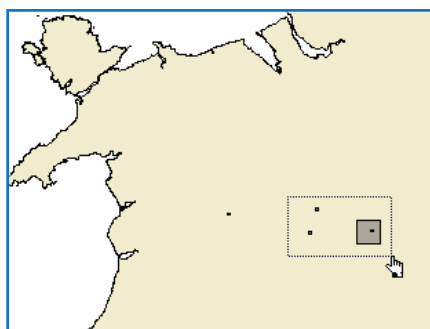
The pan tool  allows you to drag a zoomed map to see a different area.



### 7.2.2. Selecting records


The symbols representing records on the map are normally shown in blue, but selected records are shown in red.

You can draw a rectangle using the select records tool  to enclose one or more records on the map and the grid will be filtered to show just those records. This is useful, for example, to identify records of terrestrial species that fall in the sea! Turn the selection off again with the cancel button .

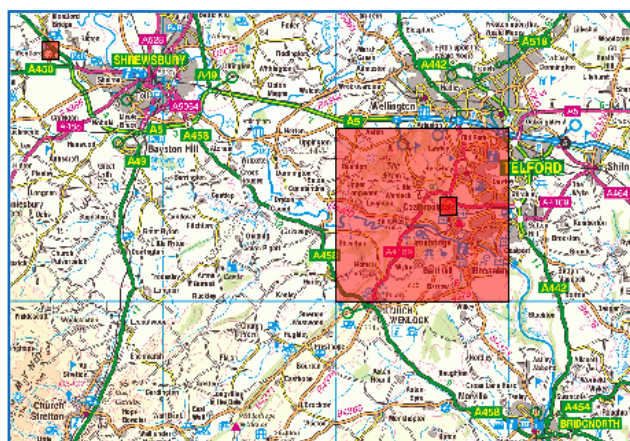


LOCATION	RECORDER	NO	GRID	SPECIES	VICE COUN
Preston Montford...	Ball, Stuart	1	SJ435145	Anasimya lineata	40
Preston Montford...	Ball, Stuart	2	SJ435145	Cheilosia variabilis	40
Earls Hill	Ball, Stuart	3	SJ4004	Cheilosia variabilis	40
Preston Montford...	Ball, Stuart	4	SJ433144	Anasimya lineata	40
Preston Montford...	Ball, Stuart	5	SJ435145	Cheilosia variabilis	40
Preston Montford...	Ball, Stuart	11	SJ435145	Anasimya lineata	40
Preston Montford...	Ball, Stuart	12	SJ435145	Anasimya transfu...	40
Preston Montford...	Ball, Stuart	13	SJ435145	Cheilosia variabilis	40


You can double-click on a spatial reference in a row of the grid and the map will redraw so that it is centred on that record. There is a balance to be struck here between zooming in far enough to be able to see which record is selected, but not zooming in so close that there are no features shown on the map which allow you to recognise where you are.

This is where the Show WMS layer button  comes in. Turning this on uses a mapping web service to fetch an OS background map which provides the necessary context. (You will see a dialog asking you to agree to the licensing terms for use of these maps).

This is helpful in indicating the location of a record, but it can also be rather slow – so you probably won't want it turned on unless you really need to find out where a record is.

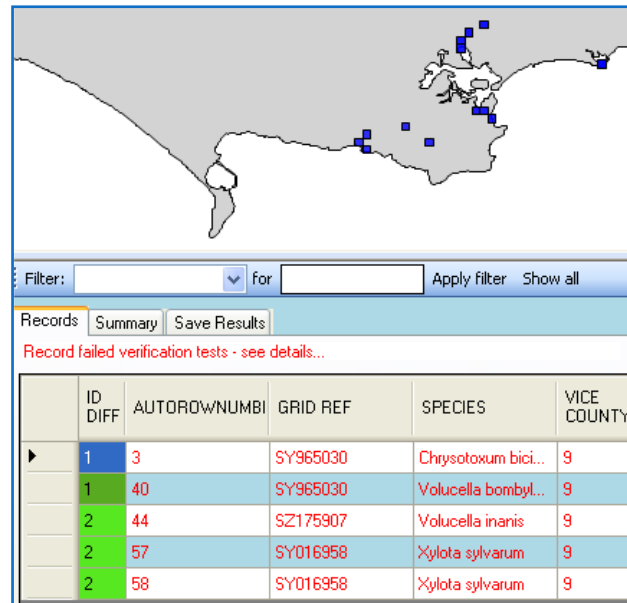



This facility is not configured by default (because we do not know of a publicly available, free WMS mapping service). If you have access to a suitable WMS map server, either run by your organisation or as a paid-for service, you will need to add the necessary details to the NBNRecordCleaner.exe.config file (see 13.1.1).

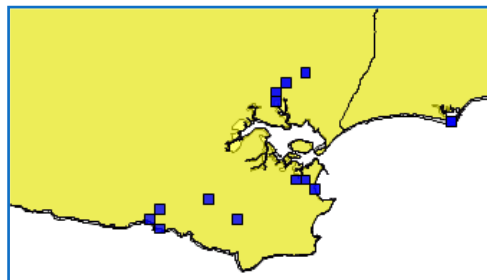
You can also add your own layers to the map using the  button to load a layer. Layers can be read from ESRI .shp files and suitably georeferenced image formats (in EPSG 27700 coordinates).


For example, it can be useful to show vice-county boundaries.

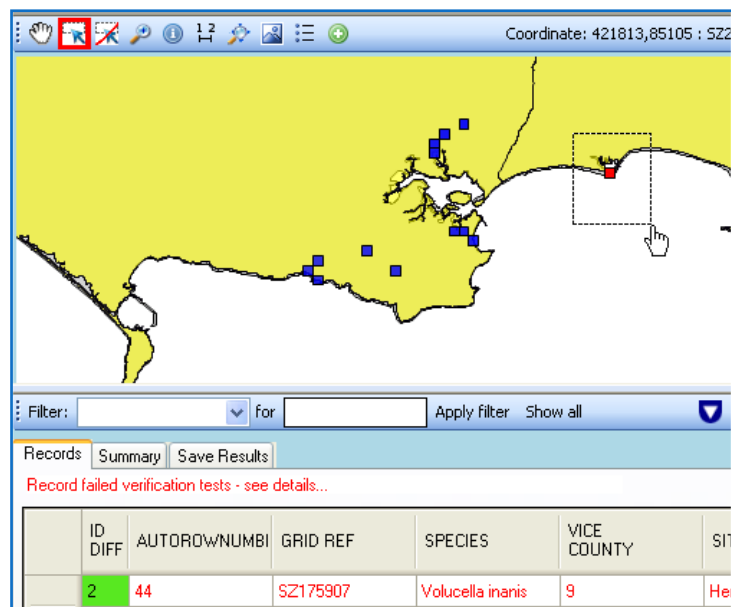
1. Here, we have run a “Coordinates in Vice-county” check and are viewing (in the Records tab) the records that have failed. We suspect that some records may have the wrong vice county assigned. Showing the vice-county boundaries on the map might be helpful.



- Click the  button, open \Record Cleaner\VerificationData\vc.shp and the map will be redrawn with VC boundaries overlain:



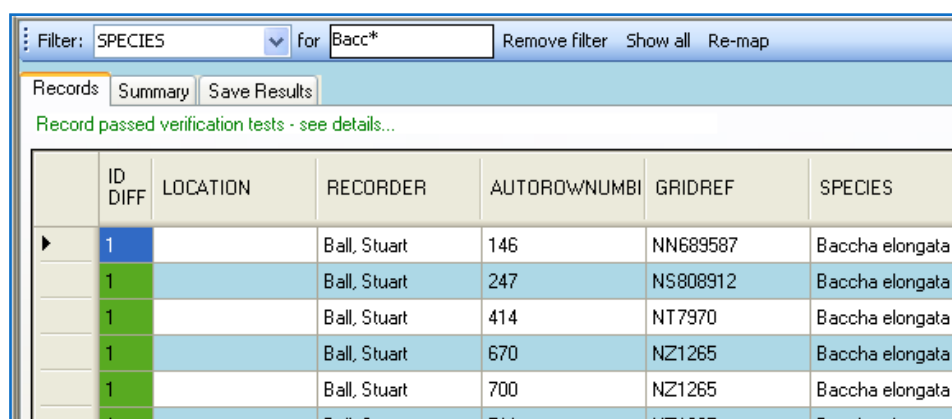
- One of these records is well over the border of VC 9. We can use the  tool to identify it.



- Now we can see that it is record 44 that has the wrong VC number – it should be 11.

### 7.3. Filtering

You can also look at (and map) a subset of the data by applying filters.



Filter: SPECIES for Bacc\* Remove filter Show all Re-map

Records Summary Save Results

Record passed verification tests - see details...

	ID DIFF	LOCATION	RECORDER	AUTOROWNUMBI	GRIDREF	SPECIES
▶	1		Ball, Stuart	146	NN689587	Baccha elongata
	1		Ball, Stuart	247	NS808912	Baccha elongata
	1		Ball, Stuart	414	NT7970	Baccha elongata
	1		Ball, Stuart	670	NZ1265	Baccha elongata
	1		Ball, Stuart	700	NZ1265	Baccha elongata
	1		Ball, Stuart	711	NZ1265	Baccha elongata

For example, to view just the records of species of *Baccha*:

1. Select “SPECIES” as the filter type in the Filter: box
2. Enter the term to search for in the “for” box. The asterisk, “\*”, is a wild card, so “bacc\*” searches for species names starting “bacc” (not case sensitive)
3. Click “Apply filter”

The grid will be redisplayed showing only those records that match the filter condition. You can now:

- “Remove filter” will remove this filter
- “Show all” will redraw the grid so that all records are visible again
- “Re-map” will redraw the map showing just the filtered records (e.g. those resulting from the filter).

Filtering is one reason to include additional data columns in the original data which are not used in validation or verification. For example, we imported a “Recorder” column, so we have the possibility to filter by recorder’s name. If this shows that most of the errors we find are in records from Fred Bloggs, we might decide that Mr Bloggs is not a reliable recorder and discard all of his records!

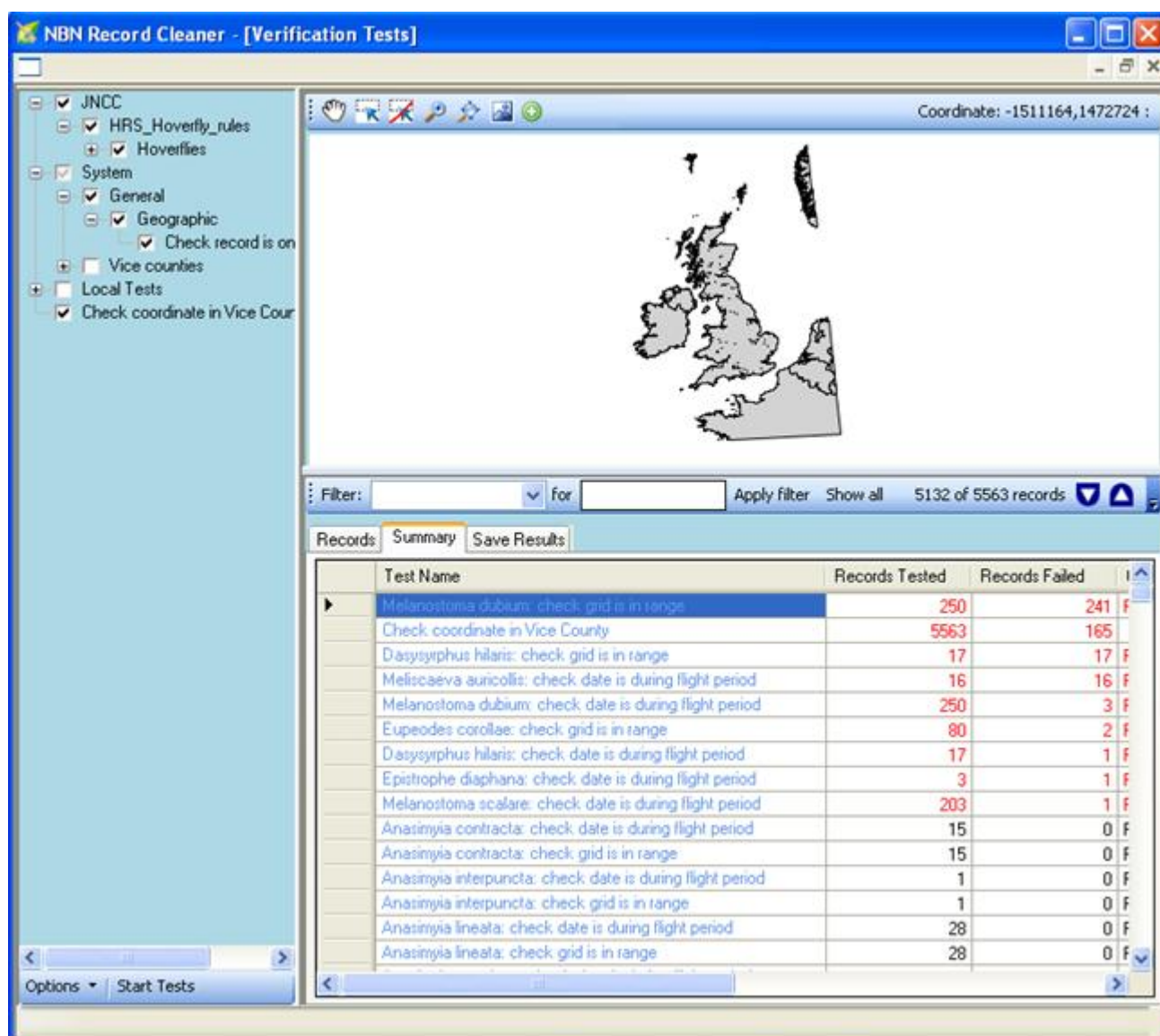
### 7.4. Running verification rules

You can select rules to apply to the dataset by ticking the boxes for the rules you want to use in the tree of available rules on the left-hand side of the screen. Choose as many rules as you want and then click the [Start Test] button.

Only rules which could be applied to the records that you have loaded are shown. In our example, we have records of hoverflies loaded, so rules relating to species in other groups will not be shown in the list of available rules.

If a filter is active when you click the [Start Tests] button, the chosen verification rules will be applied to the filtered sub-set of records only.





Here, the rules to test the date of records against known flight dates and the spatial reference against the known range of Hoverflies have been selected.

Click the [Start Tests] button, below the list of available rules, to apply the selected tests. A progress bar is displayed as the tests proceed.

**Note** that the “Check coordinates in Vice County” test takes some time. For example, for a test dataset with 500,000 records, it took around 12 minutes to complete this check.

When the checks have been completed, results are shown in the Summary tab. There is a row for each test which shows how many records that test applied to and how many of then failed. The numbers are shown in red if there were failures. For example:

- All 5,563 records were tested for coordinate in Vice County and 165 failed.
- The dates of 16 records of *Meliscaeva auricollis* were tested and all 16 were found to be outside its known flight period!
- All 15 records of *Anasimyia contracta* were both in the known range and during the known flight period. These rows are shown in black and the number of “Records Failed” is 0.

To see the records that failed a particular test, click on the “Test Name” entry and the Records tab will be displayed filtered so that only the failed records are shown. For example, click on “Check coordinates in Vice County” and the display shows the records that failed this check. Clicking “Re-map” at this point will redraw the map showing only these records.

Record failed verification tests - see details...

	ID DIFF	LOCATION	RECORDER	AUTOROWNUMBI	GRIDREF	SPECIES	VICE COUNTY
▶	1		Ball, Stuart	247	NS808912	Baccha elongata	87
	2		Ball, Stuart	248	NS808912	Melanostoma sca...	87
	2		Ball, Stuart	249	NS808912	Platycheirus albi...	87
	3		Ball, Stuart	250	NS808912	Sphegina clunipes	87
	1		Ball, Stuart	1065	NZ1557	Episyrphus baltea...	67

## 7.5. Saving results

Go to the “Save Results” tab and fill in the form. Five file formats are available from the Save as type drop down box on the Save As form:

- TAB file – tab delimited text file
- CSV File – comma delimited text file
- Excel File = Excel worksheet
- XML Recordset – text file in XML format
- Recorder Filter – a file formatted as an external filter for *Recorder 6*- i.e. a file with a .ref extension which can be opened using Tools – External Filters (see **Error! Reference source not found.**).

You can choose to export standard or additional fields available on clicking the advanced box. Move the required fields over to the fields to export box. The order of these fields can be changed by dragging the names in this box to the order you wish.

Select the export category (failed validation records, failed verification records or passed verification records) from the appropriate drop down box. These options are available depending on whether any records have passed or failed the validation or verification rules. Following an export, you will find a sub-folder has been created which contains a number of files, depending on the options you selected.

Filter: [ ] for [ ] Apply filter Show all Re-map 5114 of 5117 records

Records Summary Save Results

Select fields to include in your report:

Available fields ☒ Advanced

Coordinates of grid square  
Easting (derived)  
End date (derived)  
Northing (derived)  
Position precision  
Preferred taxon version key  
Spatial reference system (EPSG)  
Start date (derived)  
Taxon version key  
Vague Date

Add ->  
<- Remove

Fields to export (drag to reorder)

Taxon  
AutoRowNumber  
Date  
Gridref  
Identification difficulty  
Error description (validation)

Validation: Export failures

Verification: [ ]

Export all failures  
Export passed records  
☐ Add text qualifier "

Export



For example, the file “Failed Verification [Check coordinate in Vice County].CSV” resulting from the check shown above started with the following lines:

```
"TAXON VERSION KEY","SPECIES","IDENTIFICATION DIFFICULTY","DATE FROM","DATE TO","VAGUE
DATE","LOCATION","RECORDER","NO","GRID","VICE COUNTY","VERBOSE ERROR"
"NBNSYS0000007001","Cheilosia variabilis","1","31/05/2000","31/05/2000","31/05/2000","Lower
Creason Meadow","Ball, Stuart","10","SX605897","3","Coordinate not in Vice County"
"NBNSYS0000007001","Cheilosia variabilis","1","01/05/2004","31/05/2004","May 2004","Workman's
Wood","Ball, Stuart","21","SO9013","34","Coordinate not in Vice County"
"NBNSYS0000006862","Baccha elongata","1","04/06/2005","04/06/2005","04/06/2005","Southey
Woods","Ball, Stuart","31","TL1002","32","Coordinate not in Vice County"
"NBNSYS0000006862","Baccha elongata","1","04/06/2006","04/06/2006","04/06/2006","Gorthy
Wood","Ball, Stuart","38","NN957251","87","Coordinate not in Vice County".
```

## 8. Managing verification rule files

### 8.1. How rules are supplied

The flow of information from an expert group to your copy of Record Cleaner is as follows:

1. Experts create rules for the species and areas they know about. Rules are expressed in small text files, one for each rule applied to each species – so there may be lots of them! Related sets of rules (e.g. flight periods for dragonflies in GB) are zipped together for efficient download.
2. These zip files are placed somewhere on the internet that is publicly accessible (e.g. in a directory on the Recording Scheme or Society’s web-site) and registered with the NBN.
3. The NBN maintains a list of all registered rule-sets which is kept up-to-date via an automated process which is run regularly so that any changes made by the original authors are “noticed”.
4. When you start Record Cleaner (providing it finds a connection to the internet!), it gets the latest copy of the list of registered rule-sets from NBN. This is a simple text file and only needs to be downloaded if it has changed since the last time the application checked, so this check should be quick.
5. The application shows you a selection screen entitled “Verification Rules”, built from the information in the latest version of this list.
6. You choose which rule-sets you want to download and install by ticking boxes against items listed in the Verification Rules screen.
7. Record Cleaner downloads the rule-sets you select and installs the individual rule files on your computer ready for use. This step may take a while. Rule-sets could potentially contain thousands of individual rule-files (e.g. rules for all vascular plants or fungi). Consequently, the zip file containing them would be large and take a while to download, unzip and install.

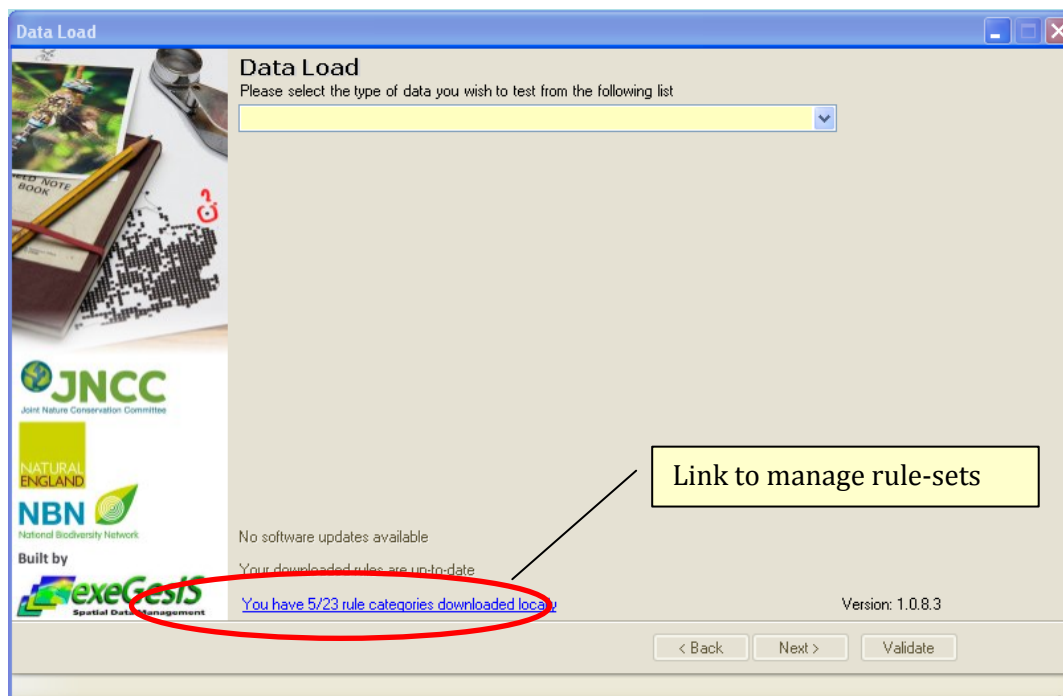
### 8.2. How rules are updated

The centrally maintained list of rule-sets includes the date on which a rule-set was last updated. Record Cleaner checks these dates at start up to see whether any of the rule-sets you have downloaded have been updated. It informs you when updates are available by showing a link “[Get updates for your downloaded rules](#)” on the initial screen. Otherwise you will see the message “Your downloaded rules are up-to-date” (not a link).

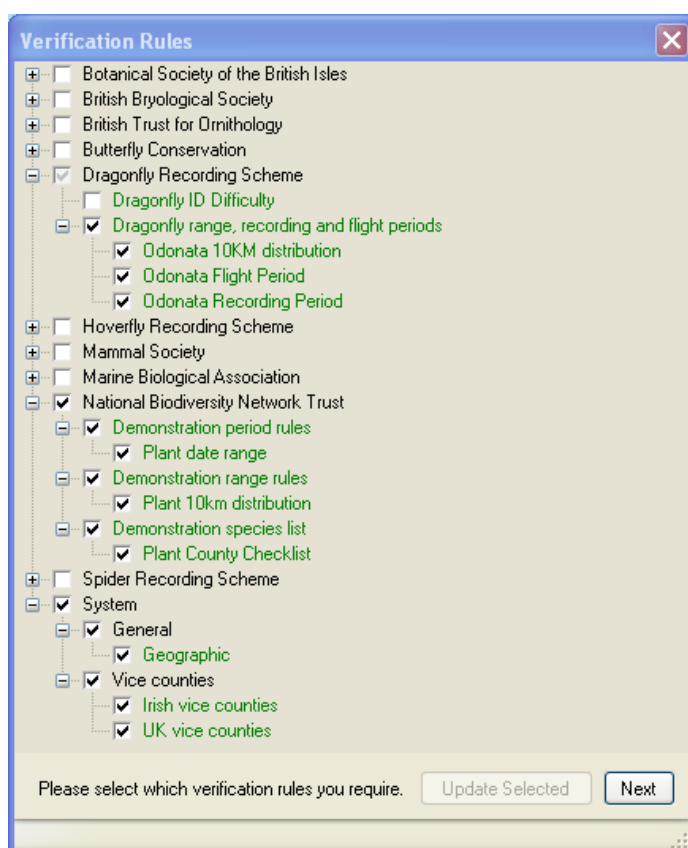
If you click the update link, Record Cleaner downloads the updated zip file(s) and replaces the rule files on your machine with those from the latest copy. This might take a few minutes. Progress is shown at the bottom of the screen.

### 8.3. Choosing rule-sets

On the initial start-up screen, you will see a link showing how many of the available rule sets you have chosen to install. When you first use the application, only one rule-set ("System"), which is included by default in the installation, will be installed.



Clicking this link opens the Verification Rules screen:



Here you see a list of the available rule-sets arranged in a hierarchy with the source at the top level (e.g. “Dragonfly Recording Scheme”) followed by the rule-sets managed by that source (“Dragonfly Recording Scheme – Dragonfly range, recording and flight periods”).

You choose which verification rules you want to install by ticking the check-boxes next to them. Once one or more categories have been ticked, the [Update Selected] button will become active. Click this button to start the process of downloading and installing the selected rule-sets. This may take some time.

Note that colour coding is used to indicate the status of rule sets:

- **Green text** indicates that the rule-set is installed locally and is up-to-date,
- **Red text** indicates that the rule-set is installed locally, but a more recent version is available (i.e. it needs updating),
- **Grey text** indicates that the rule-set is not installed locally

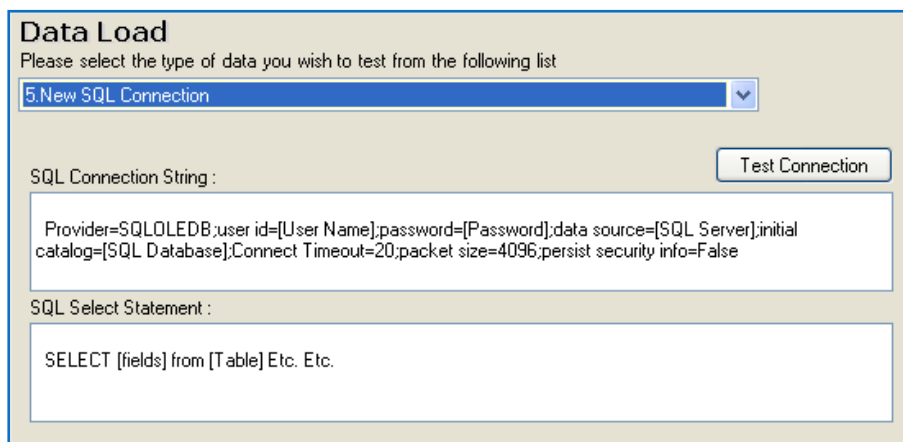
If Record Cleaner is unable to get a connection to the internet you will not be able to select new rule sets and the [Update Selected] button cannot be activated. You cannot get new rules or perform updates without an internet connection!

## 9. Loading data from other databases

This facility allows you to check any database that can be accessed by supplying a connection string. The records you want to check are obtained by running an SQL statement that you supply. This should produce a result-set containing the records you want checked with the necessary columns. You will then need to go through the column matching screens telling Record Cleaner from which of the columns in the result-set it should read the various pieces of data it needs. When this has all been set up successfully, you can save a template (which will store all of the connection string, SQL statement and column matches). You can then check the database whenever you want just by selecting your template from the list of data types in the initial screen.

As an example, we will set up a template to check data from *Recorder 2002*:

1. Run Record Cleaner and select “5 New SQL connection” from the list of data types. You will see:



2. We need to enter a connection string in the first edit box, “SQL Connection String”, which will allow us to connect to the \Recorder 2002\Database\nbndata.mdb file (Access 97 format) of

our *Recorder 2002* installation. A quick Google search reveals the connection string format to connect to an Access 97 database:

```
Provider=Microsoft.Jet.OLEDB.4.0;Data Source=<path>.mdb;User  
Id=admin;Password=<password>;
```

So our connection string will look something like:

```
Provider=Microsoft.Jet.OLEDB.4.0;Data Source=C:\Program Files\Recorder  
2002\Database\nbndata.mdb;User Id=admin;Password=;
```

Note:

- Your path may be different – this is the path for a default installation of *Recorder 2002*,
- If you are a single user, the password will probably be blank as shown here, but you will need to include the database password if one has been set on your system.

3. We also need to enter an SQL statement into the second edit box, “SQL Select Statement”. This needs to return a result-set that contains the rows and columns we want to check. You can build a suitable SQL statement using the interactive query builder in Access (or any other SQL tool you prefer). Here is a possible SQL statement

```
SELECT TAXON_OCCURRENCE.TAXON_OCCURRENCE_KEY,  
TAXON_LIST_ITEM.TAXON_VERSION_KEY, SAMPLE.VAGUE_DATE_START,  
SAMPLE.VAGUE_DATE_END, SAMPLE.VAGUE_DATE_TYPE, SAMPLE.SPATIAL_REF,  
Min(ADMIN_AREA.SHORT_CODE) AS VC  
FROM (SAMPLE INNER JOIN ((TAXON_OCCURRENCE INNER JOIN TAXON_DETERMINATION ON  
TAXON_OCCURRENCE.TAXON_OCCURRENCE_KEY =  
TAXON_DETERMINATION.TAXON_OCCURRENCE_KEY) INNER JOIN TAXON_LIST_ITEM ON  
TAXON_DETERMINATION.TAXON_LIST_ITEM_KEY =  
TAXON_LIST_ITEM.TAXON_LIST_ITEM_KEY) ON SAMPLE.SAMPLE_KEY =  
TAXON_OCCURRENCE.SAMPLE_KEY) LEFT JOIN (LOCATION_ADMIN_AREAS LEFT JOIN  
ADMIN_AREA ON LOCATION_ADMIN_AREAS.ADMIN_AREA_KEY =  
ADMIN_AREA.ADMIN_AREA_KEY) ON SAMPLE.LOCATION_KEY =  
LOCATION_ADMIN_AREAS.LOCATION_KEY  
WHERE (((TAXON_DETERMINATION.PREFERRED)=True) AND  
((TAXON_OCCURRENCE.VERIFIED)<>1) AND ((TAXON_OCCURRENCE.CHECKED)=True) AND  
((SAMPLE.SPATIAL_REF_SYSTEM) In ("OSGB","OSNI")) AND  
((ADMIN_AREA.ADMIN_TYPE_KEY) Is Null Or (ADMIN_AREA.ADMIN_TYPE_KEY) In  
("NBNSYS00000000032","NBNSYS00000000036")))  
GROUP BY TAXON_OCCURRENCE.TAXON_OCCURRENCE_KEY,  
TAXON_LIST_ITEM.TAXON_VERSION_KEY, SAMPLE.VAGUE_DATE_START,  
SAMPLE.VAGUE_DATE_END, SAMPLE.VAGUE_DATE_TYPE, SAMPLE.SPATIAL_REF;
```

Points to note:

- This is filtered so that unchecked records and records that have failed *Recorder's* internal verification are excluded

```
(TAXON_OCCURRENCE.VERIFIED)<>1)AND (TAXON_OCCURRENCE.CHECKED)=True)
```

- Only records with GB or Irish grid references are included

```
SAMPLE.SPATIAL_REF_SYSTEM In ("OSGB","OSNI")
```

- The Vice-county is obtained by the outer joins to LOCATION\_ADMIN\_AREA and ADMIN\_AREA. But there could be more than one VC given for a record. If we did not do something about it, this could result in two or more rows in the result-set with the same

TAXON\_OCCURRENCE\_KEY. In turn, this would lead to “duplicate key” validation errors from Record Cleaner. Therefore the whole thing is a GROUP BY query and the select clause “Min (ADMIN\_AREA.SHORT\_CODE) AS VC” is used so that only one VC number (the smallest) is returned per row in such cases.

4. Having set this up, we can use the [Test Connection] button to check that it works. If no error messages are reported, press [Next >] to proceed to column matching.
5. The following column matches are needed:

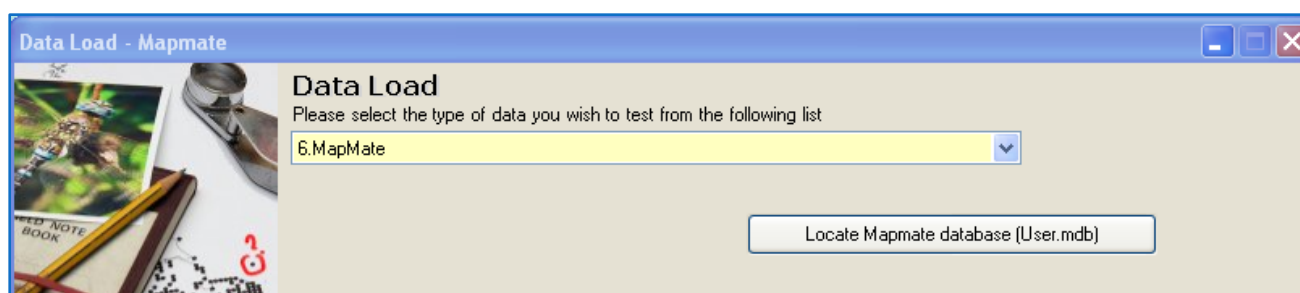
Matching screen	Field	Value
Unique Record Key Field	Unique Record Key	TAXON_OCCURRENCE_KEY
Date Fields (advanced settings)	Start Date	VAGUE_DATE_START
	End Date	VAGUE_DATE_END
	Vague date type	VAGUE_DATE_TYPE
Coordinates Fields	Actual Coordinates Type	British, Irish or CI gridref
	Enter Gridref or Easting	SPATIAL_REF
Species Field	Taxon Version Key	TAXON_VERSION_KEY
Vice County Field	Vice County	VC

If we save the settings as a template called “Recorder 2002”, validation and verification of the records from *Recorder 2002* should then be possible in future just by selecting this template from the initial screen.

## 10. Use with MapMate

The Record Cleaner needs to be run on a machine with access to the MapMate database (User.mdb). On a default installation, this is in the \MapMate\Users\ directory.

1. In the initial “Data Load” screen, select “MapMate” from the data type drop-down list.



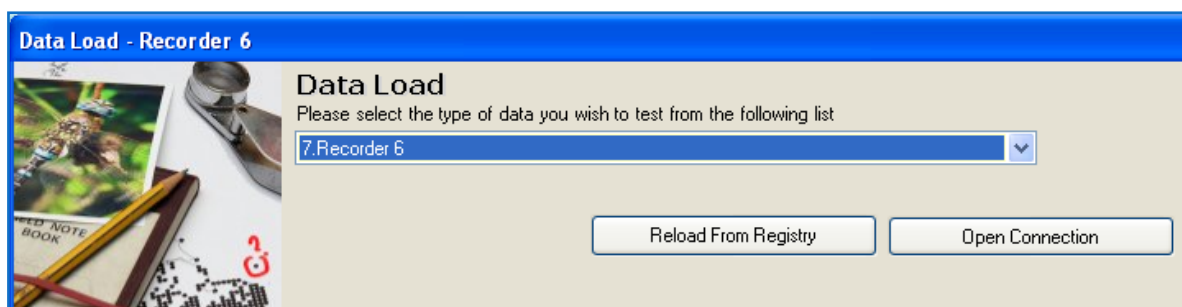
2. Click the [Locate MapMate database (User.mdb)] button and open the User.mdb file.
3. After a short pause as it reads the data from the Access database, Record Cleaner will start the validation process. There is no need to match columns to data types – Record Cleaner already “knows” which columns contain which data.
4. Proceed through the validation and verification screens as normal, running whichever checks you wish.

Note that you may get validation errors because not all the species names used by MapMate will be recognised. This is particularly likely for names suffixed with “agg.”, “sensu lat.” and “sensu strict”.

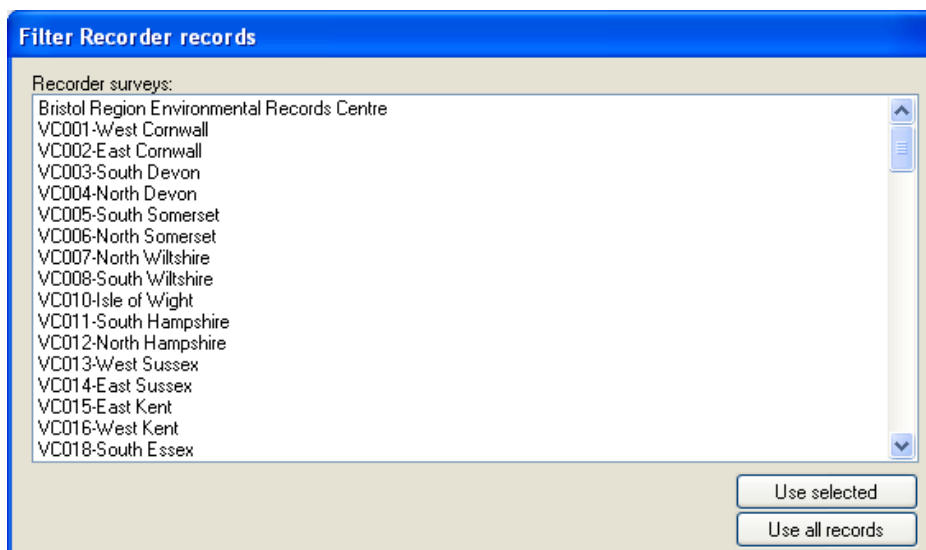
## 11. Use with Recorder 6

The NBN Record Cleaner needs to be run on a machine which has access to the SQL Server database (NBNDData\_Data.MDF ). If you run Record Cleaner on a machine where *Recorder 6* is installed it should have the necessary access.

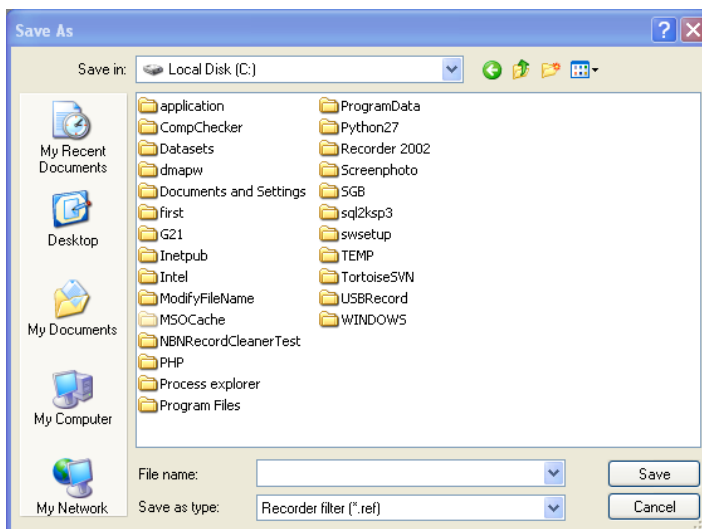
1. In the initial “Data Load” screen, select “Recorder 6” from the data type drop-down list.



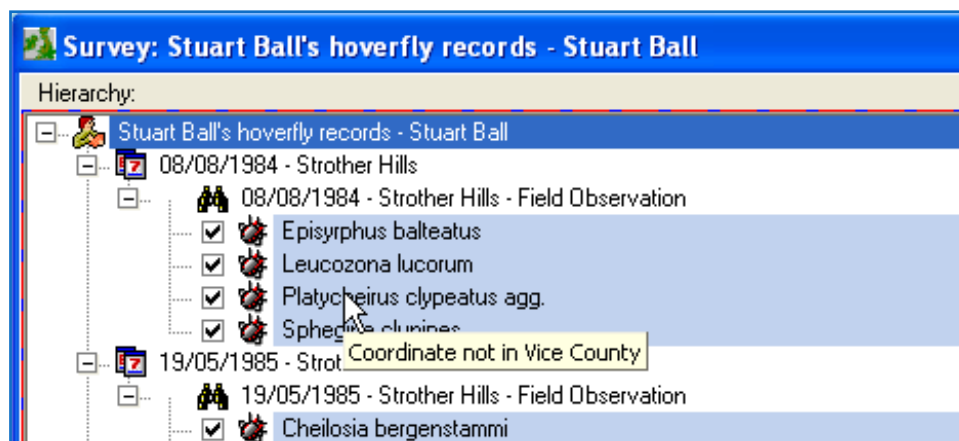
2. The first time you use Record Cleaner with *Recorder 6*, click the [Reload From Registry] button. This will read the necessary connection details from the computer’s registry. It should not be necessary to do this again unless the configuration of *Recorder 6* is changed.
3. Click the [Open Connection] button.



4. A list of Surveys will be displayed. If you want to check the entire database, click the [Use all records] button. If you only want to check certain Surveys, select them by clicking in the list and then click the [Use selected] button.
5. After a short pause as it reads the data from the SQL Server database, Record Cleaner will start the validation process. There is no need to match columns to data types – Record Cleaner already “knows” which columns contain which data.
6. Proceed through the validation and verification screens as normal, running whichever checks you wish.
7. There is a useful feature to save results to *Recorder 6* .ref files from the Save Results tab.



8. If you open *Recorder 6* and load one of these files (Tools – Load External Filter), the Observation window will open with the records detailed in the file selected. When the window opens, you will just see the name of one or more Surveys listed. Select one and then press the asterisk key [\*] **on the numeric keypad** (this is a generic Windows shortcut which can be used to expand any tree view). This will expand the hierarchy to show the records selected in the External filter file. The items with an error will be highlighted in blue and, if you hover your mouse over one of the highlighted items, the error message for that record will be shown as a tool tip.

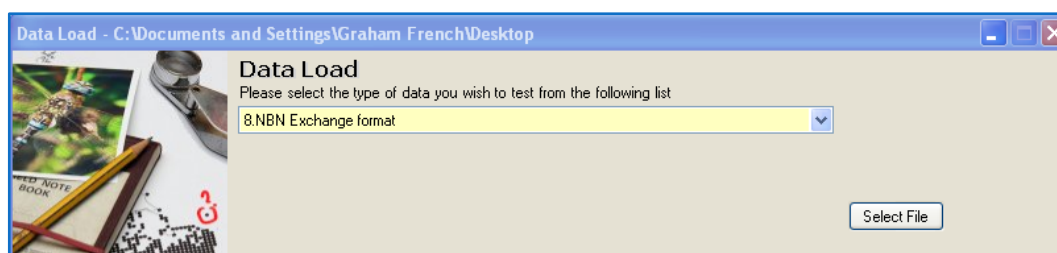


## 12. Use with NBN Exchange Format

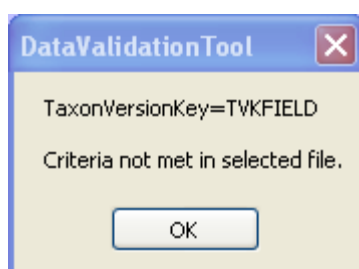
The NBN Exchange format is a text-based format which in its simplest form encapsulates the basic components of a species record (*what* was recorded, *where* it was recorded, *when* it was recorded and *who* recorded it). Consisting of a number of compulsory and optional fields, as well as some additional constraints, it is the format required for submitting datasets to the [NBN Gateway](#). Submitting your records in this format will greatly speed up the loading process, reducing any delays in getting your data onto the NBN Gateway. Further information can be found on the [NBN website](#).

1. In the initial “Data Load” screen, select “NBN Exchange format” from the data type drop-down list.





2. After selecting your dataset in tab delimited text file format there is no need to match columns to data types – Record Cleaner already “knows” which columns contain which data.
3. Record Cleaner checks that the required fields are present in your dataset. If any are missing then a message will appear with the name of the missing field(s) and Record Cleaner will not proceed any further.



If one or more required fields are missing **go back to your text file, add the missing field(s)** and then reload your text file and try again.

4. When all the required fields are present Record Cleaner will start the validation process with additional checks specified within the NBN Exchange Format (version 2.4). These additional validation checks are applied as follows:

Column name (not case sensitive)	REQUIRED/OPTIONAL	Validation checks
RecordKey	REQUIRED	30 characters or less
SurveyKey	OPTIONAL	30 characters or less
SampleKey	OPTIONAL	30 characters or less
TaxonVersionKey	REQUIRED	Must be in the species dictionary
BiotopeKey	OPTIONAL	
ZeroAbundance	OPTIONAL	Value must be “T” (true) or “F” (false)
Sensitive	OPTIONAL	Value must be “T” (true) or “F” (false)
Date	Either Date and/or the combination of the other 3 fields is REQUIRED	The year part must have 4 digits
StartDate		The year part must have 4 digits
EndDate		The year part must have 4 digits
DateType		Must be one of “D”, “DD”, “O”, “OO”, “Y”, “YY”, “-Y”, “P”, “ND”, “U”
SiteKey	OPTIONAL	30 characters or less
SiteName	OPTIONAL	100 characters or less



GridReference	Either GridReference and/or the combination of the other 2 fields is REQUIRED	Must not contain spaces
East		
North		
Projection	REQUIRED	Must be "OSGB", "OSNI", "OSI", "WGS84" or "OSGB36"
Precision	REQUIRED	
Recorder	OPTIONAL	140 characters or less
Determiner	OPTIONAL	140 characters or less
<user defined field> (attribute field)	OPTIONAL	255 characters or less

Note that the case of the column names is not significant - i.e. "Date", "date" or "DATE" would be equally acceptable. Further validation rules concerning dates and spatial references are enforced. These items can be specified in two ways - either using a single column or using a combination of 3 and 2 columns respectively. You can include columns for both the simple and combination-of-columns methods in the same dataset and use whichever is appropriate for a given record. For example, if terrestrial records (which usually employ grid references for the spatial reference) and marine records (which often use lat/long coordinates) occur in the same dataset, you will very likely want to have both a GridReference and East, North columns. **But** you cannot use both methods in the same record. So the following rows are acceptable within the same dataset:

Date	StartDate	EndDate	DateType	GridReference	East	North
27/03/2009					-4.28933	55.8785
	01/03/2009	31/03/2009	0	NS568674		

These rows are not and will be reported as validation errors:

Date	StartDate	EndDate	DateType	GridReference	East	North
27/03/2009	01/03/2009	31/03/2009	0			
				NS568674	-4.28933	55.8785

5. **Correct any records failing the validation checks in your text file.** Reload and try again. This time all the records should pass validation, including the additional checks required for the NBN Exchange Format.
6. Continue on to the verification step to map your dataset and apply any necessary verification checks. **Correct any records failing the verification checks in your text file.**

Submit your NBN Exchange Format compliant dataset, as the tab-delimited text file, to the NBN Gateway team at [data@nbn.org.uk](mailto:data@nbn.org.uk). Any further information you require about this process is provided in the [data providers pack](#). If you do not find the information required in the pack then send any further questions to the NBN Gateway team at [data@nbn.org.uk](mailto:data@nbn.org.uk).

## 13. Appendix

### 13.1. Installation directories

After a default installation on Windows XP all files are located under the C:\Program Files\NBNRecordCleaner folder.

This has sub-folders:

- UserSettings – this contains saved import templates, and lists of species that are matched for each import (so can be automatically re-applied)
- ValidationData – contains definitions for additional validation rules. At present, only additional rules for NBN Exchange format are supported.
- VerificationData – contains downloaded verification rules. After installation, it will contain two subdirectories “SystemRules” and “Personal”. SystemRules contains geographical information used for testing grid refs are on land, in vice-counties, etc. Personal is the place to put your own rule files either for customisation or for rule file development and testing.

The location of all these files can be controlled by editing the NBNRecordCleaner.exe.config file.

#### 13.1.1. NBNRecordCleaner.exe.config

NBNRecordCleaner.exe.config contains some additional settings:

- WMSUserName
- WMSPassword
- WMSURL

Which are used to specify access to a WMS service that can be displayed on the map

In addition if you access the internet through a proxy – you may need to contact your network administrator for the correct settings to enter in:

- ProxyUsername
- ProxyPassword

The PassColour and FailColour settings configure the background colour for the tool tips that popup in the verification screen reporting the results of verification tests run on a given row. The values are R,G,B (integers in the range 0-255).

```
<userSettings>
  <Record Cleaner.My.MySettings>
    <setting name="UserSettings" serializeAs="String">
      <value>C:\Record Cleaner\UserSettings</value>
    </setting>
    <setting name="ShapeOutlineFile" serializeAs="String">
      <value>C:\Record Cleaner\ShapeOutline\GB.shp</value>
    </setting>
    <setting name="ProxyUserName" serializeAs="String">
      <value />
    </setting>
    <setting name="ProxyPassword" serializeAs="String">
      <value />
    </setting>
    <setting name="WMSUserName" serializeAs="String">
      <value>xyz</value>
    </setting>
    <setting name="WMSPassword" serializeAs="String">
```

```

        <value>abc</value>
    </setting>
    <setting name="WMSURL" serializeAs="String">
        <value>http://www.securewms.no-ip.net/Map</value>
    </setting>
    <setting name="ReportsFolder" serializeAs="String">
        <value>C:\Record Cleaner\Reports</value>
    </setting>
    <setting name="PassColour" serializeAs="String">
        <value>212,254,212</value>
    </setting>
    <setting name="FailColour" serializeAs="String">
        <value>255,216,207</value>
    </setting>
</Record Cleaner.My.MySettings>
</userSettings>
<applicationSettings>
    <Record Cleaner.My.MySettings>
        <setting name="VerificationTests" serializeAs="String">
            <value>C:\Record Cleaner\TestData</value>
        </setting>
    </Record Cleaner.My.MySettings>
</applicationSettings>

```

There are also settings to alter the background colour for the five identification difficulties groups during verification

### 13.2. Data loading and validation error messages

The following table shows the error messages that may be displayed whilst loading data or as a result of data validation.

Error message	Notes
The file you have selected does not contain the same fields as your template - please check and try again.	File that you are trying to load has changed field names from when it was saved. The easiest thing to do is probably to recreate template from scratch and remove the old one.
There was an error in your data file at about line: xyz	This could well be due to a stray quotation mark. The text importer can't handle a single, unmatched quotation mark ( ' or " ) in the file. Check the row number given for error.
The spreadsheet / database you have selected does not contain the same fields as your template - please check and try again.	File that you are trying to load has changed field names from when it was saved. The easiest thing to do is probably to recreate template from scratch and remove the old one.
<b>Unique keys</b>	
Duplicate Key	Unique key field was not unique!
<b>Dates:</b>	
Not a valid date without separators	Date may not have had a delimiter (e.g. /) separating its parts.
Not a valid vague date	Tried to interpret the string as vague date but failed!
Not a valid date	As it says !

Date is in the future	As it says !
Start and end dates are not the same	Vague date of type "D" had differing start and end dates.
End date is before start date	As it says!
Start date must be 1st of month and end date last day of same month	Vague date of type "O" breaking the rules.
Start date must be 1st of Jan and end date 31st Dec of the same year	Vague date of type "Y" breaking the rules.
Start date must be 1st of Jan and end date 31st Dec	Vague date of type "YY" breaking the rules.
Start date must be 1st of Jan and end date null	Vague date of type "Y-" breaking the rules.
Start date must be null	Vague date of type "-Y" breaking the rules.
Year must be 9999 and dates from 1 to last day of same month	Vague date of type "M" breaking the rules.
Start date must be 1st of month and end date last day of a month with a year of 9999	Vague date of type "S" breaking the rules.
Not a valid season	Vague date of type "S" breaking the rules.
Unrecognised vague date type	The values given in the Vague Date Type is not one of the valid options. It is not a valid vague date.
<b>Coordinates:</b>	
Unknown Coordinate Type	Completely unrecognised coordinate type!
No grid reference supplied	String was blank
Invalid British Grid format	Invalid format for specified grid ref
Invalid Irish Grid format	Invalid format for specified grid ref
Invalid Longitude	Comma delimited Long/Lat not deciphered into 2 parts
Invalid Latitude	Comma delimited Long/Lat not deciphered into 2 parts
Long Lat is invalid - should be numeric separated by comma	Comma delimited Long/Lat not deciphered into 2 parts
Invalid UTM format	Invalid format for specified format
Invalid JTM format	Invalid format for specified format (JTM = Jersey Transverse Mercator)
Invalid UK, Irish, CI grid format	Was supposed to be a grid ref – but couldn't decode it to any available format
Unrecognised coordinate type	Gave up because the format was not recognised
Invalid Longitude format	Couldn't be decoded to a Long. It should recognise most formats (decimal degrees +/- E/W, degrees, mins, degrees mins secs etc).
Invalid Latitude format	Couldn't be decoded to a Lat
Longitude out of range	Must be -180 to +180
Latitude out of range	Must be -90 to +90
<grid type> - X out of range	Must be 6 figures or less. <grid type> can be UK grid, Irish grid or JTM.

<grid type> - X is not a number	Must be numeric
<grid type> - Y out of range	Must be less than 1300000
<grid type> - Y is not a number	Must be numeric
Invalid UTM format - cannot be supplied as X Y	Need zone etc. to be valid
Invalid UK grid - cannot be supplied as X Y	Grid refs cannot be specified as x,y in a single field. Separate into 2 fields.
Grid ref not available	Tried to decode as a grid ref, but failed
xyz unknown projection	An unrecognised projection system identifier has been encountered.
Precision does not match GridReference	If a both a grid reference and a precision are supplied, then they must match. For example, "TL123456" is a 100m square, so the precision must be 100.
5 km Precision does not match GridReference	Unlike tetrads, there is no specific format supported for giving 5km grid squares. This can be done by giving a 1km square reference (TL1234) for the bottom left corner of the 5km square and a precision of 5000. This error will occur if the precision is 5000, but the grid reference is not a 1km square.
<b>Species:</b>	
Unknown Scientific Name	Scientific name was not found in the dictionary – check and correct in Data cleansing form.
Non Unique Scientific Name	More than one entry for this scientific name was found in the dictionary. You can pick the correct one from the drop down in Data cleansing form.
Unknown TVK	Taxon Version Key was not in dictionary. You can correct by picking a species from the drop down in Data cleansing form.
Unknown Species or TVK	Catch all error – probably doesn't get shown
Common Name not matching PTVK	The Preferred Taxon Version Key for the common name is different from the PTVK for species / common name. Either correct in the data cleansing form, or re-import leaving out the common name.
Unknown Common Name	Common name was not in dictionary. NOTE: This is a warning NOT an error
Preferred tvk: <PTVK > was not found in the species dictionary	This is a dictionary error (all PTVKs must also exist as a TVK). Users shouldn't see this error unless they have been editing the species dictionary. It is meant to help dictionary editors track down errors.
Duplicate species name found in species dictionary : Name = <name>,<authority>	Users shouldn't see this error unless they have been editing the species dictionary. It is meant to help dictionary editors track down errors.
Errors Found loading species dictionary Do you wish to see them ? "	Users shouldn't see this error unless they have been editing the species dictionary – helps dictionary editors track down errors.

<b>VC</b>	
Unknown Vice County	Vice county not in the list – correct your data and re-import (or at a pinch add entry to ViceCounty.txt file).
<b>NBN Exchange format</b>	
RecordKey/SampleKey/SurveyKey/SiteKey must not be more than 30 characters	These fields are limited to a maximum of 30 characters on the NBN Gateway
ZeroAbundance/Sensitive must be T or F	The possible values for these fields are True or False which are specified as “T” or “F”.
Year in Date/StartDate/EndDate must be 4-digits	You must give 4-digits for the year (i.e. 01/11/10 is not acceptable, must be 01/11/2010).
Date in both StartDate and Date field	Dates in an NBN Exchange Format file can be supplied in either a single date field or in three fields, StartDate, EndDate and DateType. But you can only use one of these formats in any one data row.
Sitename must not be more than 100 characters	The site name field is limited to 100 characters on the NBN Gateway
Grid Reference must not contain spaces	As it says! (The NBN Gateway data loading system is more strict than some other systems)
Projection must be OSGB or OSNI or OSI or WGS84 or OSGB36	The NBN Gateway only accepts these projection systems for spatial references.
Spatial reference in both Gridreference and East fields	Spatial references in an NBN Exchange Format file can be supplied in either a single GridReference field or in two fields, East and North, containing coordinates. But you can only use one of these formats in any one data row.
Recorder/Determiner must be less than 140 characters	These fields are limited to a maximum of 140 characters on the NBN Gateway
Field must not be more than 255 characters	User defined fields are limited to a maximum of 255 characters on the NBN Gateway

### 13.3. Reinstalling on Windows 7 machine

You may need to uninstall and re-install NBN Record Cleaner on your machine if issues arise with the use of NBN Record Cleaner or the downloaded verification rules. When using the application on a Windows 7 machine an additional step may be required during the uninstalling process. Information on the step is provided below.

Installation of the NBN Record Cleaner using the default path should occur in 2 places

1. **C:\Program Files (x86)\NBNRecordCleaner** - installs the main files used by the application
2. **C:\ProgramData\NBNRecordCleaner** - stores the users settings eg downloaded rules and templates when the program is run as an administrator.

When installing and running the NBN Record Cleaner **it is recommended to always use administrator rights** (right click NBN Record Cleaner icon and run as administrator). This ensures that

any user settings are written to the ProgramData\NBNRecordCleaner file. If not then the user settings are written to a third user folder

### 3. C:\Users\UserName\appData\local\VirtualStore\ProgramData\NBNRecordCleaner.

This can result in different versions of the rules and templates being written to different folders which may or may not be used depending on whether the application is run with administrator rights or not. This can lead to problems eg rules and templates apparently disappearing, species dictionary not updating correctly.

When uninstalling the application this additional user folder is not deleted so following uninstallation check that the above 3 folders are deleted before re-installing (as an administrator!). The 2 ProgramData folders are hidden and so may not appear on your machine by default. To unhide the folders go to **Control Panel - Folder Options - View Tab** and tick **show hidden files, folders and drives**

Following re-installation run the software updates by clicking on the link in the data load form. This will download and update the latest species dictionary used by the application.

By doing this fresh install most of the issues with using the NBN Record Cleaner with Windows 7 should be corrected.

## 13.4. Clearing verification rules

Occasionally you may need to clear all the verification rules you have downloaded to use in the NBN Record Cleaner and start again. There is no option to do this directly within the tool. To remove all the rules requires uninstalling and then reinstalling NBN Record Cleaner on your machine. Reinstalling the software should only take a few minutes to complete but a quicker way of removing all the rules is to manually delete the rules yourself.

Following these steps should provide a quick and easy way to do this.

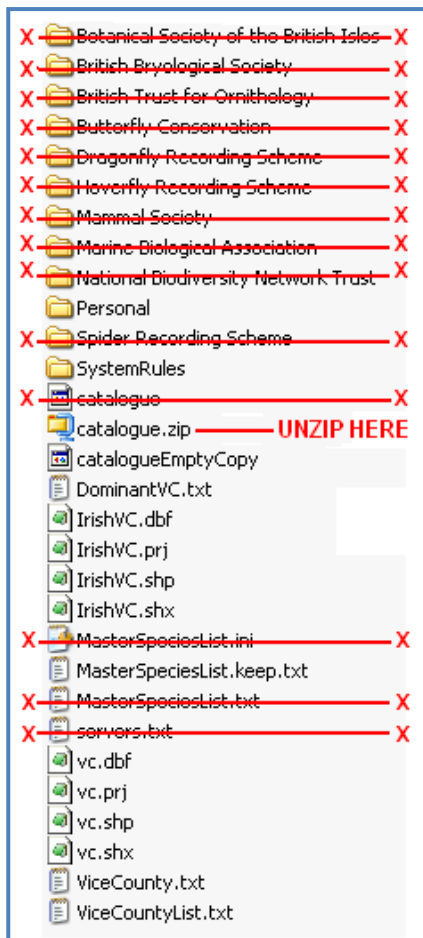
1. Close NBN Record Cleaner if open
2. Navigate to the VerificationData folder in the NBNRecordCleaner folder on your local machine. In this folder

In Windows XP the default path is C:\Program Files\NBNRecordCleaner\VerificationData

In Windows 7 the default path is C:\ProgramData\NBNRecordCleaner\VerificationData. By default this folder is hidden– to unhide the folder go to Control Panel - Folder Options - View Tab and tick show hidden files, folders and drives.

3. Delete all subfolders EXCEPT the 2 subfolders – Personal and SystemRules
4. Delete the catalogue file, replacing it with the fresh catalogue file by unzipping catalogue.zip and moving the new catalogue file into the VerificationData folder if necessary.
5. Delete the MasterSpeciesList.ini and MasterSpeciesList.txt files if present but KEEP the MasterSpeciesList.keep.txt file
6. Delete the servers.txt file





If you are using NBN Record Cleaner on Windows 7 then there may be additional VerificationData folder stored in the Users directory. Verification rules are stored here if the application is not run as an administrator. Check for the presence of this folder at C:\Users\UserName\appData\local\VirtualStore\ProgramData\NBNRecordCleaner\VerificationData

If this folder exists then

1. Delete all subfolders
2. Delete the catalogue file, replacing it with the fresh catalogue file (see step 4)
3. Delete servers.txt



The verification rules should now be cleared. Open NBNRecordCleaner and click on the rule categories download link (which should report 0 rules categories downloaded) to download any required verification rules again.

If you are having further issues then it is best to uninstall and reinstall NBN Record Cleaner on your local machine again. This will remove the downloaded verification rules.